

Sonification

Extending the Visual Arts experience

Thesis by:

Thomas Fink (6703402)

Supervisor: **Almila Akdag**

Second supervisor: **Anja Volk**



Utrecht University

Graduate School of Natural Sciences

Utrecht University

Netherlands

08-04-2022

Contents

1	Introduction	3
1.1	Problem statement	3
1.2	Research approach	4
1.2.1	Research questions	4
1.2.2	Research iterations	4
2	Preliminary knowledge	5
2.1	Sound properties	5
2.2	Music theory	5
3	Literature review	6
3.1	Sonification	6
3.1.1	First experiments on data sonification	6
3.1.2	Sonification of paintings	6
3.1.3	Color Coding Sound	6
3.1.4	Eyes-Free Art	7
3.1.5	Color information	8
3.1.6	Edge detection	10
3.1.7	To the realm of art	10
3.2	Feature extraction from visual data	12
3.2.1	Low level	12
3.2.2	High level	12
3.2.3	Scene detection	14
3.2.4	Object detection	15
3.3	Feature extraction from audio	16
3.4	Music Generation Using Deep Neural Networks	17
3.4.1	Symbolic	17
3.4.2	Non-symbolic	18
3.5	Sound programming frameworks	18
3.6	Contribution	19
4	Methodology	20
5	Model 1: The first	20
5.1	Low-level feature sonification design ideas	20
5.1.1	The use of the dominant color	20
5.1.2	Color of a segment	21
5.1.3	Edges as timbre	22
5.1.4	Histogram as timbre	25
5.1.5	Edges as melodies	26
5.1.6	Panning based on location	26
5.1.7	Navigation of segments: salience	26
5.2	Technical implementation low-level features: visual	27
5.3	Technical implementation low-level features: audio	31
5.4	High-level feature sonification	34
5.4.1	Technical implementation	34
5.5	Description of the sound	35
5.6	Intermediate results	36

6	Model 2: object segmentation and FM synthesis	36
6.1	Object segmentation from high-level features	36
6.1.1	New panning	37
6.1.2	Segments as melody	37
6.2	FM synthesis	38
6.3	Description of the sound	39
6.4	Intermediate results	40
7	Model 3: Instruments	41
7.1	Description of the sound	42
7.2	Intermediate results	42
8	Model 4: Instruments accompanied by FM synthesis	43
8.1	Inner scaling	43
8.2	Objects as an influence in note duration	43
8.3	Description of the sound	44
8.4	Intermediate results	44
9	Evaluation by experts	45
9.1	Dataset	45
9.2	Interview questions	46
9.3	Results per painting	46
9.3.1	Charge of the scots greys at waterloo	47
9.3.2	Alfred Sisley - Snow at Louveciennes	49
9.3.3	Enrique Simonet El - barbero del zoco	52
9.3.4	John Lavery - The Fairy Fountain	54
9.3.5	Claude Monet - Water Lilies	57
9.3.6	William Merritt Chase - The Olive Grove	59
9.3.7	Valentin Serov - Iphigenia in Tauris	62
9.3.8	Paul Delvaux - The Viaducto	64
9.4	Two paintings, one sonification	66
9.4.1	Charge of the scots greys at waterloo & William Merritt Chase - The Olive Grove	66
9.4.2	Valentin Serov - Iphigenia in Tauris & Claude Monet - Water Lilies	67
9.4.3	Alfred Sisley - Snow at Louveciennes & Enrique Simonet - El barbero del zoco	68
9.4.4	Paul Delvaux - the viaducto & John Lavery - The Fairy Fountain	69
9.5	General feedback	70
10	Conclusion	71
11	Discussion and future work	72
A	Appendix: Dataset filter stages graphs	77
B	Appendix: Paintings from the dataset	80

1 Introduction

For most people auditory and visual experiences are separated, but for some people these experiences are more connected[10]. People having the phenomenon “visual-ear” can have an auditory sensation while looking at visual stimuli. This auditory sensation appears to come from moving or flashing stimuli. However, it makes one wonder if this auditory sensation could also apply to static stimuli, e.g. to paintings. Enjoying paintings is mostly a visual experience, but connecting paintings with audio or music is not a far stretch[35]. Museums are always looking for ways to make art more engaging for the overall public. The transformation of visual stimuli to auditory stimuli is a known research line that especially addresses the needs of people with visual impairment, and emphasizes certain data types (e.g., information visualizations)[12]. In this research project we will focus on visual arts, especially paintings, and explore the possibilities of painting sonification to extend the art experience. Research has already been done on the sonification of paintings[5][18][25], but this research mainly focuses on using color and color properties, e.g. hue for their sonification. Other research[32] does look at visual features besides color but does not provide an automated process for this sonification. In this research we extend this sonification by using AI to extract high-level features present in paintings and create an automated process for the sonification. We argue that this leads to a sonification that is better at pleasantly conveying a painting’s content and a simpler method for generating the sonification. Besides opening up a new way to enjoy visual arts, rendering a new dimension within the exhibition space, and offer the public a deepened experience to engage with an art collection, research outcomes might be used in the future to help the visually impaired with the enjoyment of paintings. Within the framework of the museums of the 21st century that makes use of Artificial Intelligence technology usually for restoration, analysis, and re-creation of art, this approach will bring a fresh perspective on how AI and visual arts can be coupled. The research will lay out possible designs of sonification methods including high-level visual features. Two designs are proposed, namely a design including scene detection and a design including scene and object detection. All the aforementioned designs share the use of low-level visual features, namely, color and edge.

1.1 Problem statement

The enjoyment of paintings is a purely visual one. Thereby, extending this enjoyment to the auditory space was the main purpose of this research. There have been several attempts to make art experiences more accessible for the visually impaired[39], e.g. by converting paintings to textures. Another methods is the sonification of visual input. The field of data sonification for visual data tries to represent data in the auditory space instead of representing it in a visual manner, which is hard or impossible to use for people with visual impairment. This idea of data sonification has been applied to paintings. However, current research on the sonification of paintings[32][5] does not provide an automated process for a pleasant sonification with the incorporation of high-level visual features present in paintings. This research, therefore, aims to extend the visual art experience by automating and extending current sonification methods. The high-level visual features discussed in this research are scenes and objects.

1.2 Research approach

1.2.1 Research questions

From the above problem statement, we defined the following research question: “How can high-level visual features present in paintings be incorporated in an automated and pleasant painting sonification method.” From the research question, we define three subquestions:

- How can existing sonification methods contribute to the automation of painting sonification?
- How can a sonification pipeline be created to incorporate high-level features extracted from paintings?
- How will the overall quality and the value of the addition of high-level features to the sonification be validated?

The first question leads the literature research into existing sonification methods. The result of this research is a guideline for the development of an automated sonification method for paintings. The second question aims to guide the development of an extended sonification pipeline that can incorporate high-level features. The result is a framework where high-level features can be added to create a more elaborate sonification. The last question exists to create a telling evaluation of the created sonification. As the quality and fittingness of the sonification are hard to quantify, quantitative evaluation will be near to impossible. Therefore a well-designed user study is created and provides the answer to this question.

1.2.2 Research iterations

To answer the first question and create a baseline implementation, the research cycle found in Figure 1 was used. Here the feature research section in the prototyping phase will be covered by the literature research done on existing sonification methods. When the implementation is done the evaluation phase will start. If the results of the evaluation analyses are not satisfactory the prototyping phase will begin again. If a good baseline has been created, the same research cycle will be used for the addition of high-level features. Because of the additive nature of the proposed design, this research cycle will be carried out per the addition of a feature.

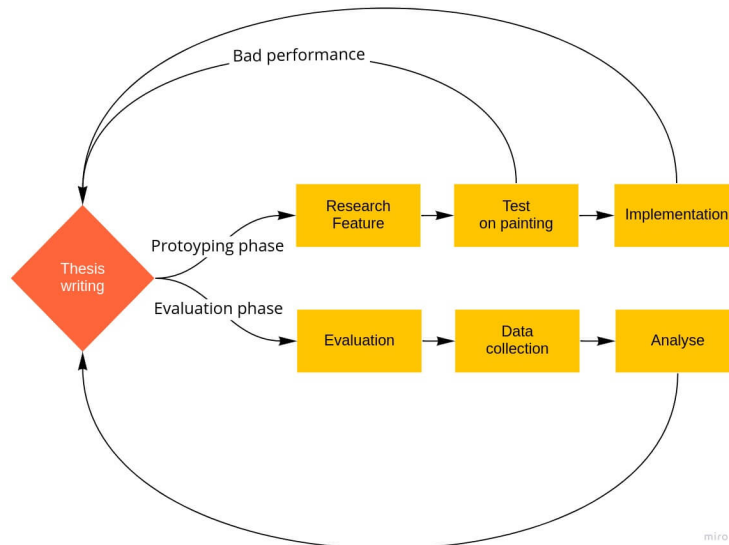


Figure 1: Research cycle

2 Preliminary knowledge

In this section, short explanations will be given on terms that are needed to get a better understanding of the research.

2.1 Sound properties

Sound property	Meaning
Pitch	Frequency of a sound wave. Higher frequency means higher pitch
Loudness	Perceived loudness of a sound, often notated in Decibel
Timbre	Character of the sound. The same note played on a different instrument can sound very different

2.2 Music theory

Music property	Meaning
Note	A single sound with a specific pitch and duration. 12 notes exist in western music. I.g., C C# D D# E F F# G G# A A# B
Melody	Sequence of notes to create a musically pleasing pattern
Chord	Multiple notes, at least three, played at the same time creating a musically pleasing sound together
Chord progression	A sequence of chords sounding musically pleasing
Octave	The frequency multiple or half of a note. E.g., middle C has a frequency of 262Hz. C with one octave higher has a frequency of 524Hz and with one octave lower is 131Hz
Consonant	Notes sounding pleasant together
Dissonant	Notes sounding unpleasant together

3 Literature review

3.1 Sonification

To find the best way to sonify paintings, existing work in the field of data sonification is researched and compared to create a base for this research. Within the field of sonification, there are two directions[12], namely high-level and low-level sonification. In high-level sonification, e.g., text to speech, symbolic data is used to transform information into the auditory space, whereas for low-level sonification, low-level visual data is used, e.g. color information. Within this research, the focus is on the latter, as the visual information of paintings is used to steer the sonification process.

3.1.1 First experiments on data sonification

One of the first researches that has been done on representing information within the auditory space has been carried out by Pollack and Ficks[28]. They tried to convey binary states based upon different auditory stimuli. The goal of this exploratory research is to find out if it is possible to encode such binary states within an auditory display. Therefore, they created an audio display containing an alternation between a tone and a noise. The sound properties of the tone and noise are used to encode the binary information by linking the states to eight sound stimuli: frequency of the noise, loudness of the noise, frequency of the tone, loudness of the tone, the rate of alternation, the on-time fraction of the tone, the total time of presentation, and the direction of the sound within the room. By directly linking a binary state to a specific change in audio, e.g., the frequency could be high or low, or the sound could be loud or quiet, they showed that people are capable of extracting this information from the auditory space. As this research was of exploratory nature, no ideas for practical applications were mentioned. One application is the sonification of visual data. People with a visual impairment are unfortunately faced with the impossibility of obtaining information via visual means. Therefore the idea of transforming visual information to the auditory space has been used to aid them in obtaining such visual information[43][4][29].

3.1.2 Sonification of paintings

The idea of using sonification to convey visual information has also been applied to the field of visual arts. One direction is to help people with visual impairments gain access to visual arts more easily with respect to paintings. Most existing visual sonification methods are not designed with paintings in mind, as they transform visual data directly into the auditory space with a one-on-one mapping to stay as close to the source data as possible[43][4]. This is effective to convey information accurately but easily results in a non-musical sonification. While for some artistic purposes this method can be desirable[18], for most people, a pleasing sonification for art would be more appreciated. Therefore, research has been done to create a more pleasing sonification for paintings[32][25][5].

3.1.3 Color Coding Sound

Cho et al.[5] took the approach of using color as a base for the sonification. To accomplish this they selected a specific instrument to play for a certain color. This color-instrument mapping is mostly inspired by the findings of the painter Wassily Kandinsky. For example, red is played by a violin while green is played by a cello. Besides color, the color intensity was represented by a specific chord or melody. This linking of a specific sound to a specific selection of colors, they called sound color coding. The researchers proposed three different color coding models. The first model directly linked chords and instruments to colors where a different color is played by a different instrument. The saturation of the color maps to the intensity of the sound and the lightness maps to the pitch of the chord. The second model used the same principle as the first model, but linked

the saturation and intensity to different melodies from Vivaldi’s “Four seasons”. Saturated sound uses Vivaldi’s “Spring” composition, light uses “Autumn” and dark is based on “Summer”. The third model is based on classical music scores where instruments are again linked to colors, whereas saturation and color intensity is linked to a specific part of the classical score best representing those values. Saturation has an intense and clear melody, light uses high and fast notes and dark creates more separation between the notes. The last two models were recorded in collaboration with a sound designer, composer, and performers. By using these color-coded sounds of the third model as a base, they made a manual composition of the Starry Night painting of Van Gogh. The painting was formed into a piece of sheet music by taking vertical partitions and playing them left to right, see Figure 5.

3.1.4 Eyes-Free Art

While Cho et al.[5] only focused on transforming color information within paintings into the auditory space in a musically pleasing manner, Rector et al.[32] incorporated more features in their representation of a painting in the auditory space by creating Eyes-Free Art. Users could select background music, sonification, sound effects, or a verbal description by standing in a designated area, see Figure 4. The background music is added to convey the mood and genre of the viewed artwork, which is selected by a survey done via [Amazon Mechanical Turk](#). The sonification method is used to give the user a pleasing auditory experience. For their sonification method, they used an orchestral loop and changed the loudness of different instruments to convey color, e.g., bright green would produce a loud piano, see Figure 2. The painting was manually divided into segments and those segments got assigned a color by hand, see Figure 3. To give a more literal impression of the painting, sound effects are added manually for objects within the painting. The verbal description was a manually curated description of the painting, describing information about the painting as well as describing its aesthetics. The users were able to navigate the painting by moving one hand in the air. Tracking of hand position and location was done using a Microsoft Kinect. Rector et al. tried to convey more context of a painting by separately including background music, sound effects, and a verbal description. However, the sonifications created by Rector et al. are composed by hand. They state that automating parts of their process is future work. Rector et al. used an orchestral loop for their sonification of color information as they considered existing work of generative methods for the sonification of color, such as the later-described work of Cavaco et al.[4], were musically unpleasant. However, a generative approach better suits the automated goal of this research, Therefore, generative approaches will be discussed further.

Color (RGB)	Orchestra	Piano	Harp
Red (255, 0, 0)	100	10	10
Purple (255, 0, 255)	100	10	100
Blue (0, 0, 255)	10	10	100
Teal (0, 255, 255)	10	100	100
Green (0, 255, 0)	10	100	10
Yellow (255, 255, 0)	100	100	10
White (255, 255, 255)	100	100	100
Gray (128, 128, 128)	50	50	50
Black (0, 0, 0)	10	10	10

Figure 2: Mapping of color to the loudness of instruments[32]



Figure 3: Colored segments used for sonification[32]

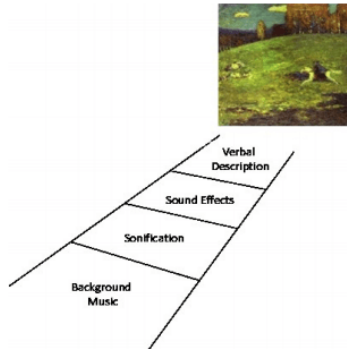


Figure 4: Selectable areas[32]





Partition	Color and Music Composition
Night sky with stars and moon	 <p>Blue/S/Cello (Bach: Cello Suite No.1 in G) Yellow/S,L/Trumpet (Haydn: Trumpet Concerto In E-Flat Major, Hob.VIle:1—I. Allegro)</p>
Night sky and whirlwind	 <p>Blue/S/Cello (Bach: Cello Suite No.1 in G) Green/D/Oboe (Rossini Variations for oboe) Yellow/S/Trumpet (Haydn: Trumpet Concerto In E-Flat Major, Hob.VIle:1—I. Allegro)</p>
Cypress trees and whirlwinds	 <p>Blue/S/Cello (Bach: Cello Suite No.1 in G) Green/D/Oboe (Rossini Variations for oboe) Yellow/S,L/Trumpet (Haydn: Trumpet Concerto In E-Flat Major, Hob.VIle:1—I. Allegro)</p>
Cypress trees and cold land	 <p>Blue/S,D/Cello (Bach: Cello Suite No.1 in G) Green/D/Oboe (Rossini Variations for oboe)</p>

Figure 5: Composition of Starry Night[5]. S stands for saturated, L stands for light and D stands for dark.

3.1.5 Color information

To help the visually impaired to understand color within a picture a tool has been created by Cavaco et al.[4]. This tool extracts color information from videos or still images and converts it into the auditory space. The color attributes that are used for the conversion are hue (e.g., blue, red, magenta, etc.), saturation (colorfulness, e.g., deep blue or pale blue), and value (the intensity of the color), see Figure 7. These variables are linked to the psycho-acoustic variables of sound, namely pitch, timbre, and loudness. Where the hue is linked to the pitch, saturation influences the shape of the waveform to affect timbre, and value changes the loudness. The picture is scanned from top to bottom in a vertical manner where each row is played sequentially, as can be seen in Figure 6. A row is divided into 12 segments, with each segment having a shifted phase waveform based on the x-axis, which plays simultaneously. To validate the tool a user study was done with eight visually impaired participants. The experiment was a forced-choice test where sonification was carried out on one of seven colors. The participants noted that the frequency of neighboring colors was too close together and thereby hard to distinguish. When taking the accuracy of correctly chosen colors or their neighbors, meaning it would be correct if the purple was the ground truth and the neighboring color violet was chosen, the accuracy was 86.25 percent. While the tool can accurately convey color information present in visual input, the tool directly maps color information to sound. A downside of one-to-one mapping is that the resulting audio has no musical relation, and is, therefore, less suitable for a pleasant art experience. Polo et al.[29] tried to address this problem by the addition of harmony.

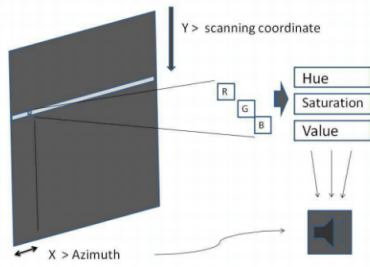


Figure 6: The architecture of the sonification tool by Cavaco et al.[4]

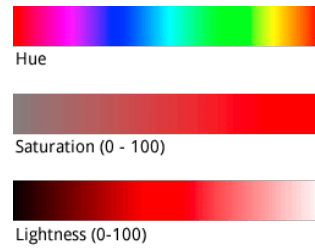


Figure 7: Hue, situation, and lightness in numeric space

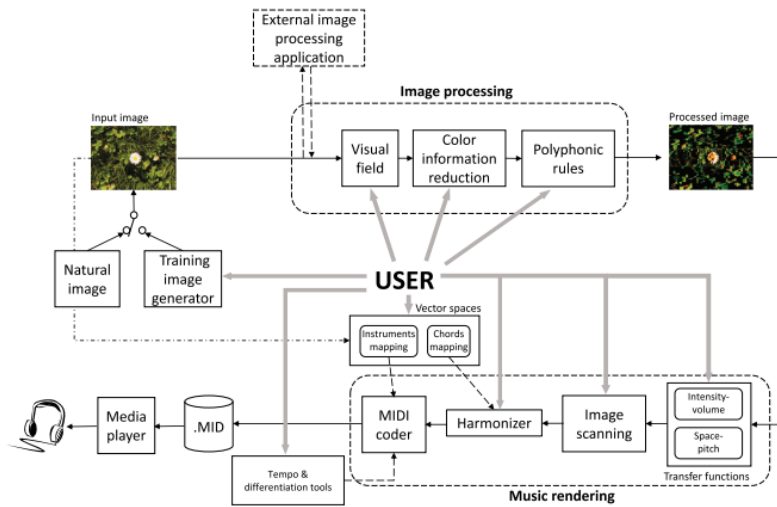


Figure 8: Framework architecture: Musical Vision[29]

Polo et al.[29] created a tool called Musical Vision. The tool took the same approach of using color as a feature for the sonification of visual input. However, they used the RGB (Red, Green, Blue) values instead of hue, saturation, and intensity. Their reasoning is that RGB values closer represent the cone structure found in the eye. The RGB values represent three notes that play simultaneously, one for red, one for green, and one for blue. For image processing, they tried to stay closer to the biological process of human vision. This was done by emulating the retina where central and peripheral vision has a different resolution. This means the center of the image will contain more detail and the surroundings are decimated. The part of the image that serves as the center of vision was defined by the user. Besides the resolution reduction, color information within the image is also reduced. The reason for this reduction originates from the idea that the visual spectrum can carry more information than the auditory space. Another difference to previous work is making the sonification harmonious, meaning multiple notes play at the same time in a musically pleasing manner. The work also stands out by giving the user control over the image processing and music rendering, as can be seen in Figure 8. To evaluate the performance, a user study with 12 participants was conducted. A participant would listen to the sonification of an image and in a forced-choice test style, the participant was asked to pick the best fitting image while multiple images were being presented. With training, non-musical participants had an accuracy of 70 percent, whereas musical participants could draw by ear. This method only transforms color information into the auditory space, thereby it is hard to discern objects from the sonification.

3.1.6 Edge detection

To help the visually impaired recognize objects within a picture, Yoshida et al.[43] created the tool EdgeSonic. To accomplish this, the authors used an edge detection algorithm to transform the original picture to edge features. The user can touch an image using a touch screen. This touched area will then be used for the sonification. If an edge is present in the touched area this edge will represent a line, as can be seen in Figure 9. This line will be used to control the pitch, meaning the frequency of a sound wave, of the produced sound. When a line goes up in a left-to-right manner, the pitch of the sound present will increase. When the line goes down, so will the pitch. If there is no up or down movement the pitch will stay the same. To help the user with navigation, a beep with a specific interval is presented. The closer the user is to an edge, the shorter the interval between beeps will become. This research mainly focused on edge detection and therefore did not include a way to convey the color information or other features of the image. To test the performance of the framework before and after training, sighted participants were blindfolded and told to explore a given shape and reproduce it. After 90 minutes of training two out of four participants were able to reproduce the shapes as can be seen in Figure 10. While the tool can accurately convey shape information present in visual input, the tool directly maps edge information to sound. Consequently the resulting audio has no musical relation and is, therefore, less suitable for a pleasant art experience. Using the sounds of those objects, as done by Rector et al.[32], could be considered a more pleasant experience.

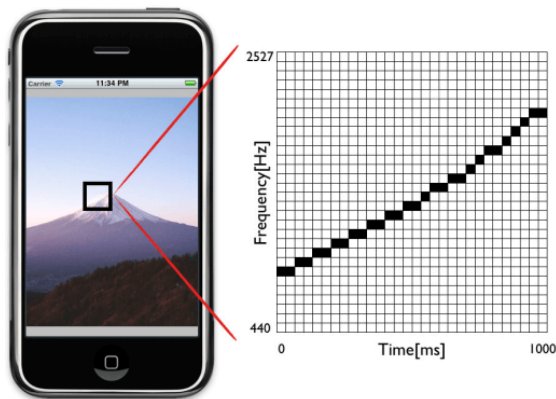


Figure 9: Conversion of edge line to a frequency[43]

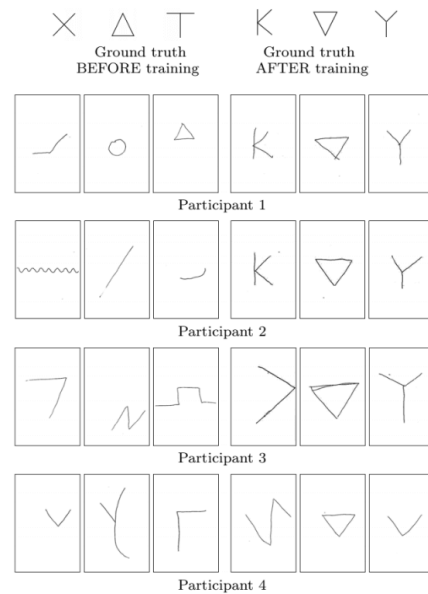


Figure 10: Results before and after training[43]

3.1.7 To the realm of art

Kabisch et al.[19] used sonification to create an art installation. This project was not meant to help people with a visual impairment but shows that sonification does not have to be a tool that directly transfers information and can be of artistic value. In this installation an image of a 360-degree landscape, see Figure 11, was projected on a circle of fabric hanging in the air. The position of users within this circle was tracked as input for the sonification along with edge detection and color information. Kabisch et al. argue that the mapping of data to sound in the realm of art is as much an artistic choice as a philosophical and technical choice. Thereby having the luxury to

interpret data in a way where it can communicate more than the source material itself. They set out to use data extracted from landscape images to influence the created sound on the macro-level (rhythm and form) as well as on the micro-level (timbre). The image obtained from edge detection is scanned on the vertical axis. When the RGB values reach a set threshold, those values are used to trigger notes with a certain pitch based on vertical position. The position of the user is used to select a horizontal area of the picture. To change the representation on a micro level the edges are used to create a wave-shaping lookup table. This table consists of different edges which can be used to shape a waveform for the audio generation, see Figure 12. To give a more literal impression of the image ambiance recorded while the photo of the landscape was taken is played in the background. This idea of using sound present at the scene is thereby in line with the addition of object sounds done by Rector et al.[32]. The addition of the scenic sound is however also done manually, therefore research will be done on the automatic extraction of such high-level visual features. A quick overview of the discussed work and their mapping of low-level features to sound can be found in Table 3.

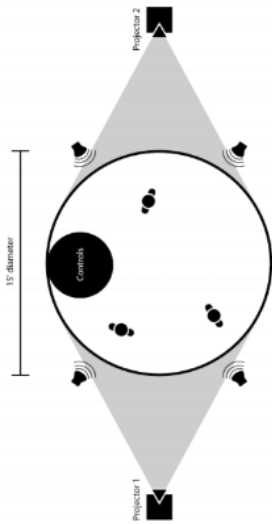


Figure 11: Top-view of installation setup[19]

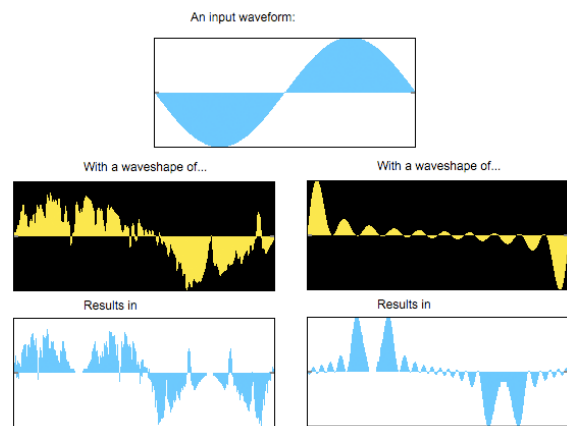


Figure 12: Waveshaping of waveform

Research	Visual features	Linked audio properties	Composition method	Navigation method	Evaluation method
Yoshida et al.[43]	Edge	Pitch	One to one mapping	Touch	User study
Cavaco et al.[4]	Color	Pitch, timbre and loudness	One to one mapping	Vertical top to bottom	User study
Polo et al.[29]	Color and position	Pitch, timbre and loudness	Harmonious mapping	User definable	User study
Kabish et al.[19]	Color, position and edge	Pitch and timbre	One to one mapping	Vertical top to bottom	No evaluation
Cho et al.[5]	Color	Pitch, timbre and loudness	Color coding	Left to right	User study
Rector et al.[32]	Color	Loudness	One to one mapping	User selectable	User study

Table 1: Sonification methods overview

3.2 Feature extraction from visual data

All visual to audio sonification methods above used extracted features of an image. In this section, we describe more in-depth information about feature extraction from visual images. This is split into two sections, low-level, and high-level features. For the high-level features, extracting of scene and object are discussed as they are included as features in the proposed designs found in section 5.4.

3.2.1 Low level

Low-level image features are image characteristics that are captured for the purpose of recognition and classification (such as pixel intensity, pixel gradient orientation, color). El et al.[13] set out to compare low-level feature extraction algorithms. These algorithms use color, edge, and corner detection to create a low-level vector space of interesting key points. These key points can for example be used for image matching, object detection, and tracking of movement. While these low-level features are not capable of extracting semantic features, they are useful for creating a map of salient points within an image, see Figure 13.



Figure 13: Salient key points found by low-level feature extraction

3.2.2 High level

High-level or semantic image features are the features commonly used by humans to describe images (e.g. objects or scenes). Low-level features do not directly correlate to semantic features. Therefore, machine learning classifiers such as Support Vector Machines (SVMs) are used to map high-level semantics to the low-level features. However, creating a direct mapping can introduce a semantic gap[42]. To combat this problem, high-level dimensional vectors, called mid-level semantics are created from the image. The features of the mid-level semantics needed to be manually selected and represented by a model such as a Bag of Words[21]. With the appearance of the Convolutional Neural Networks (CNN) the need for the creation of handcrafted feature extraction was rendered unnecessary. This is implicit in the architecture (Figure 14) of CNN models. Each layer within a CNN can learn to extract features of interest and therefore does not need handcrafted feature extraction methods. Feature extraction is done by the means of convolution. The convolution process runs a filter across the input image to extract information, these features are then passed to the next layer of the network where another convolution is done on the output of the previous layer. A convolution reduces the information present in its input by extracting only relevant information. Therefore, the resolution of the image also decreases in this process. An example of such a filter for a convolution operation can be seen in Figure 15. The model learns what filter values are needed by means of training i.e. what filter values are needed to be able to link the input to the category "cat". An example of these learned features can be seen in Figure 16, which shows that the second layer emphasizes edges, while the third emphasizes eyes. The feature maps are from shallow to deep in a left to right fashion, meaning the leftmost feature is extracted by the first convolutional layer in the network, and the rightmost feature is extracted by the last

convolutional layer of the network. This shows that the deeper a CNN is, i.e. the more layers it contains, the more abstract a feature can become. The last layer of a CNN uses the output of all the convolutions and tries to predict the correct output label for a given input.

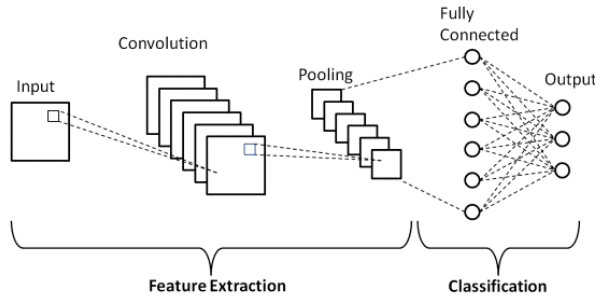


Figure 14: CNN architecture

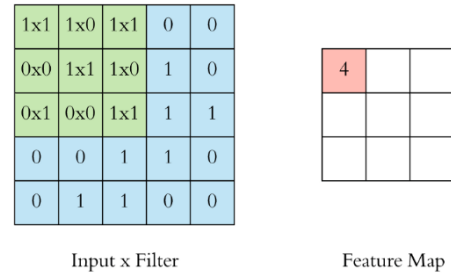


Figure 15: CNN convolution filter

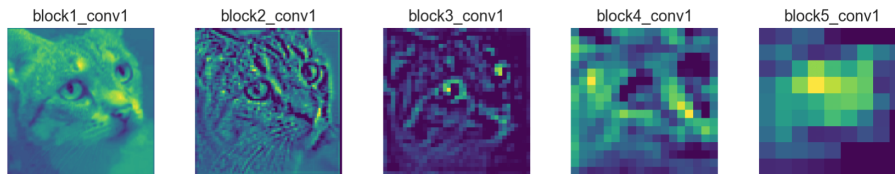


Figure 16: CNN feature maps

The depth of a CNN is of central importance to visual recognition tasks. However, deeper networks tend to exhibit degradation problems, meaning the deeper a network gets, the greater the training error becomes [17]. He et al. [17] counter this problem by the creation of Residual Networks (ResNet), one of the most successful CNNs. ResNet counteracts the problem of degradation by adding a shortcut connection over multiple layers, as can be seen in Figure 17. This shortcut contains an identity function, a function that always returns the same output as its given input (Figure 18), to add information from previous layers and prevent degradation. The upcoming sections will explore existing high-level feature extraction methods incorporated in the sonification designs found in section 5.4.

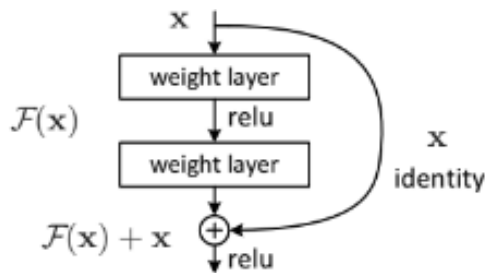


Figure 17: Residual learning shortcut [17]

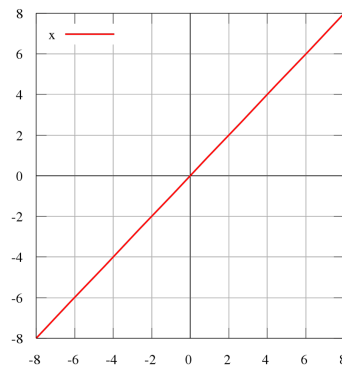


Figure 18: Identity function on real numbers

3.2.3 Scene detection

The key aspect of scene detection is to identify the place in which objects are present. For example, a table in a classroom is used for another purpose than a table in a kitchen. Most datasets have focused on object categories and less on the place of those objects. Zhao et al.,[44] therefore, created such a scene dataset and compared the performance of multiple CNNs on the subject of scene detection. The created dataset consists of around 10 million images with 434 scene semantic categories, which are the same categories in the already existing SUN(Scene Understanding)[41] dataset but with greater content. To create their dataset, they used image search engines and synonyms of categorical words to create a more extensive dataset. The ground truth is checked by using the survey framework Amazon Mechanical Turk[6], where the ground truth is verified by multiple people. To compare performance differences on smaller and larger datasets, the dataset is divided into multiple subsets. They also compared performance when CNNs were trained on already existing datasets, namely ImageNet[7] and SUN.

- Places365-Standard: 1.8 million training images, 5000-3000 images per category
- Places365-Challenge: same categories as standard but with 8 million images
- Places205: 2.5 million images from 205 categories, 15000-5000 per category
- Places88: 88 categories common with ImageNet and SUN

They compared the performance of three popular CNNs architectures. AlexNet[20], GoogLeNet[38], and VGG 16 convolutional-layer[36] and trained the aforementioned models on the Places365-Standard and Places205 datasets. Furthermore, they fine-tuned a Residual Network (ResNet)[17] on the Places365-Standard dataset. To create a baseline they trained a linear SVM (Support-Vector Machine) based on AlexNet-CNN features over 5000 images per category in Places205 and 50 images per category in SUN205, where SUN205 contains the same categories as Places205. Results of the comparison can be seen in Figure 19 and Figure 20. Within the results, they make a

	Test set of Places205		Test set of SUN205	
	Top-1 acc.	Top-5 acc.	Top-1 acc.	Top-5 acc.
ImageNet-AlexNet feature+SVM	40.80%	70.20%	49.60%	80.10%
Places205-AlexNet	50.04%	81.10%	67.52%	92.61%
Places205-GoogLeNet	55.50%	85.66%	71.6%	95.01%
Places205-VGG	58.90%	87.70%	74.6%	95.92%

Figure 19: Results of model comparison on places 205[44]

	Validation Set of Places365		Test Set of Places365	
	Top-1 acc.	Top-5 acc.	Top-1 acc.	Top-5 acc.
Places365-AlexNet	53.17%	82.89%	53.31%	82.75%
Places365-GoogLeNet	53.63%	83.88%	53.59%	84.01%
Places365-VGG	55.24%	84.91%	55.19%	85.01%
Places365-ResNet	54.74%	85.08%	54.65%	85.07%

Figure 20: Results of model comparison on places 365[44]

distinction between top-1 and top-5 accuracy. This distinction is made because the output of the model is not a single prediction. Rather, the model outputs a list of probabilities linked to scene categories. Therefore, the top-1 accuracy represents the accuracy when the prediction with the highest probability corresponds to the ground truth. The top-5 accuracy represents the accuracy of the ground truth being present within the 5 highest predicted probabilities of scene categories. This was done because scene categories can be fairly ambiguous. Keeping this in mind the top-5 accuracy shows good results and therefore can be a good addition to a sonification method to convey high-level context present in the visual data.

3.2.4 Object detection

Object detection is a well-researched subject in the field of computer vision and can be divided into three main categories: Objectness Detection (OD), Salient Object Detection (SOD), and Category Object Detection (COD)[16]. OD focuses on the general object detection of every type of object and its position. The result of OD is square proposals with the place of an object and its objectness score, see Figure 21(a). SOD aims to mimic human visual attention by highlighting the most interesting objects that should draw attention within an image, see Figure 21(b). COD tries to categorize detected objects by annotating them with a category label, see Figure 21(c). Han et al.[16] set out to explain different categories and provide a comparison between existing methods within each category. Relevant for this research is their comparison of COD models because of the semantic nature of category annotation. COD can be split into two categories, two-stage, and one-stage networks. Most older COD techniques use a two-stage network where the first stage generates proposals of boxes possibly containing objects while the other stage tries to classify the object category. In earlier work the first stage was accomplished by the use of handcrafted low-level features[13][16]. For the second stage a classifier would be used e.g. a SVM. In later work the use of CNN's created a significant improvement in the object detection performance. Firstly, CNNs were applied in the second stage of the networks to help with better classification e.g. R-CNN[15] and Fast R-CNN[14]. However, the creation of the object region proposals was still computationally slow. To improve on this problem, faster R-CNN[34] also uses a CNN network for the object region proposal stage of the network. One-stage networks are simpler in their design as they use one CNN for both object proposal and object category classification, therefore a separate proposal generation process is not needed. This simpler design trades accuracy for efficiency. Two common one-stage networks with real-time performance are YOLO[33] and SSD[22]. A comparison between the network's accuracy and speed can be seen in Figure 22. The results show good enough performance and can therefore be a good addition to a sonification proof of concept.

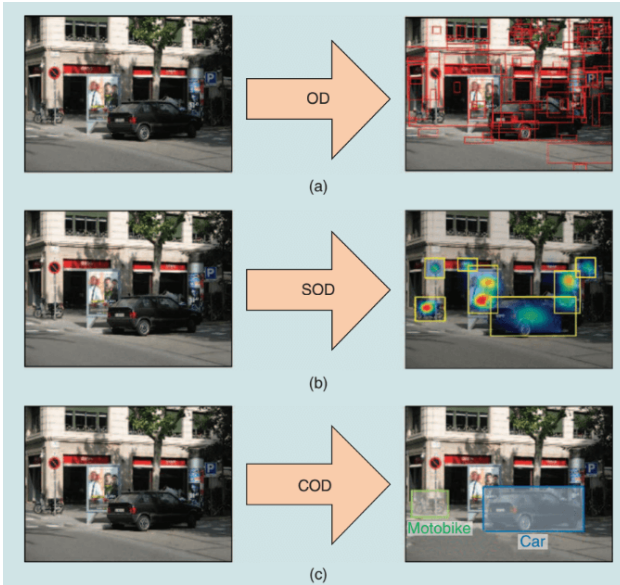


Figure 21: Three main object detection categories[16]

Method	mAP	FPS
Fast R-CNN	70.0	0.5
Faster R-CNN	73.2	7
Faster R-CNN (ResNet-101)	76.4	5
SSD300	74.3	46
SSD512	76.8	19
YOLO	63.4	45
YOLOv2 288x288	69.0	91
YOLOv2 352x352	73.7	81
YOLOv2 416x416	76.8	67
YOLOv2 480x480	77.8	59
YOLOv2 544x544	78.6	40

Figure 22: Accuracy and speed comparison on PASCAL VOC 2007 performed on Geforce GTX Titan X[16]

3.3 Feature extraction from audio

In the previous sections object and scene detection for visual input data was discussed. However, scene and object detection are not only limited to the visual space. Aytar et al.[1] tried to transfer this knowledge to the auditory space, thereby creating SoundNet. This inspired the inclusion of scene and object detection on sound within the designs found in section 5.4. SoundNet is a deep CNN that is trained to attain the ability of acoustic object/scene classification on sound. To train SoundNet a student-teacher procedure is employed to transfer discriminative visual knowledge from visual recognition models to the sound modality. This was done by employing visual recognition models on unlabeled videos and training a CNN directly on the waveform present in the audio track of the video. While there is a dependence on visual information during training, the CNN is not trained with visual data as input, and thereby no visual information is needed as input during runtime. The video data used for training is collected from Flickr and is over one year in total length. To evaluate their result they tested SoundNet on three publicly available datasets: DCASE Challenge[37], ESC-50[27], and ESC-10[27]. The DCASE dataset consists of 10 audio files in 10 categories each 30 seconds of length, making 50 minutes of total audio-of acoustic scenes and sound events. The ESC-50 dataset includes 2000 short 5-second audio clips of environmental sounds in 50 equally balanced categories, with each category containing 50 samples. The ESC-10 dataset is a subset of the ESC-50 dataset and consists of 10 classes. Results on the DCASE dataset can be found in Figure 23. For results on the ESC datasets see Figure 25. To compare normalization techniques, different teacher networks, depth of the network, the use of plane annotated audio dataset, and the use of the unlabeled videos, models with different architectures or training methods are implemented, see Figure 24 for results. The results show a higher accuracy with the use of unlabeled video for training than other state-of-the-art models at the time. The accuracy also shows that SoundNet could be useful in the design of a proof of concept sonification incorporating audio-based scene and object detection.

Method	Accuracy	Accuracy on	
		ESC-50	ESC-10
RG [29]	69%	47.8%	81.5%
LTT [21]	72%	72.9%	92.2%
RNH [30]	77%	69.5%	89.8%
Ensemble [34]	78%	71.1%	89.5%
SoundNet	88%	72.9%	92.2%

Figure 23: Result of SoundNet on the DCASE dataset[1]

Comparison of	SoundNet Model	Accuracy on	
		ESC-50	ESC-10
Loss	8 Layer, ℓ_2 Loss	47.8%	81.5%
	8 Layer, KL Loss	72.9%	92.2%
Teacher Net	8 Layer, ImageNet Only	69.5%	89.8%
	8 Layer, Places Only	71.1%	89.5%
	8 Layer, Both	72.9%	92.2%
Depth and Visual Transfer	5 Layer, Scratch Init	65.0%	82.3%
	8 Layer, Scratch Init	51.1%	75.5%
	5 Layer, Unlabeled Video	66.1%	86.8%
	8 Layer, Unlabeled Video	72.9%	92.2%

Figure 24: Comparison of SoundNet accuracy on different network depth and training[1]

Method	Accuracy on	
	ESC-50	ESC-10
SVM-MFCC [28]	39.6%	67.5%
Convolutional Autoencoder	39.9%	74.3%
Random Forest [28]	44.3%	72.7%
Piczak ConvNet [27]	64.5%	81.0%
SoundNet	74.2%	92.2%
Human Performance [28]	81.3%	95.7%

Figure 25: Result of SoundNet on the ESC datasets[1]

3.4 Music Generation Using Deep Neural Networks

Most of the discussed sonification works use a direct mapping of visual properties to sound properties for their music generation. Progress in the field of music generated by the use of deep neural networks piques interest. However, according to Briot et al.[3], there are several concerns when it comes to generating music with deep learning methods:

- The level of control
- Structure
- Creativity
- The interactivity of the process

With the above concerns the level of control seems to be most important for this research, as the output needs to be representative of the visual input. Therefore the next sections will focus on the level of control of existing work. Existing work can be split into two directions, namely models with symbolic or non-symbolic output. Symbolic output can be viewed as written down notes e.g. sheet music and does not imply the timbre of a note. Thus, symbolic models are mostly concerned with the structure of generated music. The symbolic representation of music in the computational space is mostly in the MIDI format. The non-symbolic output does not have an intermediate symbolic notation as this output is directly in the form of sound waves. Therefore non-symbolic models are mostly responsible for the way their outputs sound and not their musical structure. In the upcoming sections, the level of control of symbolic methods is discussed first, and secondly, the level of control of non-symbolic methods.

3.4.1 Symbolic

MusicVAE is a model created by Ranjan et al.[30] that uses Variational Autoencoders to generate a long-term musical structure. The Autoencoder architecture incorporates an encoding and decoding part. The encoder tries to encode input data by fitting it into a smaller dimensional latent space. Whereas the decoder must be able to recreate the input from its encoded form. The latent space can be sampled from or manipulated to create new and interesting outputs. Thereby three ways to control the output are proposed. By randomly sampling from the latent space, through interpolation of the latent space between existing musical sequences (Figure 26), and by the manipulation of existing musical sequences with attribute vectors or a latent constraint model (Figure 27). With a latent constraint model, one can define what specific part of the latent space is added or subtracted from a given input. While control over this generation seems broad the exact outcome of the model is still hard to steer. Therefore this method of music generation is not suitable for this research.



Figure 26: [Interpolation](#)

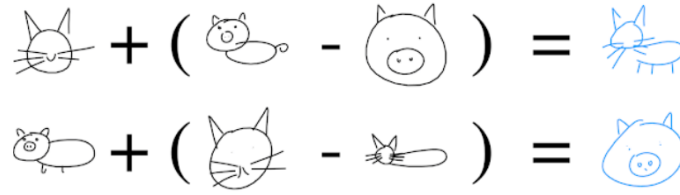


Figure 27: Latent space constraint

3.4.2 Non-symbolic

The Jukebox model by Dhariwal et al.[8] takes on the challenge of not only generating musical structure but also sound. The input to Jukebox is a small piece of an existing song and its goal is to create a continuation on the input. The architecture of Jukebox is a type of VAE, namely a Vector Quantised-Variational AutoEncoder (VQ-VAE)[31]. Therefore the network can compress the high dimensional space of audio into a lower-dimensional space. The output of the model is impressive but is not directly usable within this research. The model output is raw audio and therefore individual components cannot be separated from the output e.g. generated piano melodies, and as the output is simply a continuation of its input, the output cannot be easily controlled. Where Jukebox tried to incorporate the generation of musical structure NSynth by Engel et al.[9] only focuses on the generation of sound. NSynth is based on a WaveNet-style[26] network that learns the temporal embeddings of sounds. With this, NSynth can learn the features that make up an instrument. This ability is used to create a morphing between instruments that is more extensive than just interpolation. To be able to learn features and verify the learned feature a dataset that consists of 306043 musical notes for 1006 instruments is used[9]. Although NSynth can create interesting sounds there is no predictable way to know how the morphing of different instruments will sound and is therefore not suitable for this research.

3.5 Sound programming frameworks

The lack of control makes music generation by the use of deep neural networks not a good fit for this research. Therefore, the idea of existing sonification methods of linking visual features to audio properties will be used, see section 5. To accomplish this an audio programming framework needs to be used. There exist many programming frameworks specifically made or including audio generative methods. A couple of common frameworks are quickly compared in Table 2. As can be seen in Table 2 all frameworks offer the same general oscillators and complex wave shaping, besides Processing. PureData and MaxMSP are the two frameworks that offer visual programming, while MaxMSP is also the only framework that is not free to use. As visual programming is not suited for this research and Processing does not allow the complex waveshaping used in the design discussed in section 5, Chuck, SonicPi, and SuperCollider are considered the best options.

Name	Programming language	Oscillators	Platforms	Price
Processing	Java	Sine Saw Square Triangle Pulse	Windows, Mac OS X, and Linux	Free
PureData	Visual	Sine Saw Square Triangle Pulse Complex wave shaping	Windows, Mac OS X, and Linux	Free
MaxMSP	Visual	Sine Saw Square Triangle Complex wave shaping	Windows and Mac OS X	\$399
Chuck	C-like object-oriented language	Sine Saw Square Triangle Pulse Complex wave shaping	Windows, Mac OS X, and Linux	Free
SonicPi	Ruby	Sine Saw Square Triangle Pulse Complex wave shaping	Windows, Mac OS X, and Linux	Free
SuperCollider	C++	Sine Saw Square Triangle Pulse Complex wave shaping	Windows, Mac OS X, and Linux	Free

Table 2: Comparison of common audio programming frameworks

3.6 Contribution

Some of the above-described researches present us with generative tools which can be used to automatically transform color or edge information to audio, with a goal in mind to help people with a visual impairment. While most research focuses on accurately transferring information with one-to-one mappings[4][43], another tries to focus on user flexibility and pleasantness of the created sonification[29]. This idea has also been applied in the world for visual arts to make paintings more accessible for the visually impaired. However, the proposed methods for painting sonification are based on manual processes. This is due to the focus on creating a pleasant sonification while also introducing more high-level information into the sonification such as objects[32]. This shows that sonification can be used as a way to help people with visual impairment. This is, however not the only way sonification can be applied as Kabisch et al.[19] show by creating an art installation using sonification. This research wants to take the accomplishments of the previous research and extend the sonification realm by creating an automated system that can create an aesthetically pleasing sonification for paintings that incorporates high-level features present in a painting to extend the art experience.

4 Methodology

For this research the following research question is defined: “How can high-level visual features present in paintings be incorporated in an automated and pleasant painting sonification method.” To answer this question in a methodical manner a couple of sub-questions were created:

- How can existing sonification methods contribute to the automation of painting sonification?
- How can a sonification pipeline be created to incorporate high-level features extracted from paintings?
- How will the overall quality and the value of the addition of high-level features to the sonification be validated?

To answer the above questions a couple of different models have been created, each model consists of a visual processing component and an audio generation component. Both of these components can be divided into their respective low-level feature component and their high-level feature component. In the upcoming sections, each component of each model will be explained in detail.

5 Model 1: The first

5.1 Low-level feature sonification design ideas

Existing work on sonification has been researched to find the answer to the question: “How can existing sonification methods contribute to the automation of painting sonification?”. From this research, a low-level feature sonification design has been created. The design follows the idea of Cavaco et al.[4] of linking HSV values to sound properties but also takes inspiration from Polo et al.[29] of a noncontinuous linking of color to piano notes to create a more harmonious sound. While color influences the pitch and loudness of the generated sound, a similar approach as used by Kabisch et al.[19] will be implemented for edge information. The edge, if present in a segment, is thereby used to change the timbre of the sound with the process of waveshaping. This is with the idea that a rough sound can represent rough shapes[2]. The design ideas can be split into different steps. Each step will be described in more detail in the following sections. The technical implementation will be discussed in-depth in later sections.

5.1.1 The use of the dominant color

As a first step, the dominant color of the overall painting will be calculated to influence the scale root and scale mode of the generated musical piece. The root note of the scale will be chosen by dividing the hue space into twelve segments corresponding to the twelve keys present on a piano for one octave, see Figure 30. The mode of the scale is chosen by the lightness of the extracted dominant color. The only mode possibilities will be the Major and Minor scale, where Major will represent light paintings and Minor dark paintings. The use of the dominant color has a specific reason. Instead of taking the mean color, the assumption is that humans do not extract the mean color from a painting while looking at it. Therefore the dominant color is used to be more in line with what is represented in the painting, see Figure 28.



Figure 28: The extracted mean and dominant color of a painting by Arthur Streeton. The middle square represents the mean color of the painting of the left, while the right square represents the dominant color.

5.1.2 Color of a segment

Visual information can carry more information than is possible to represent in the auditory space[29]. Therefore, an input painting will be divided into square segments, see Figure 29. To reduce the visual information available, only the dominant color of a segment will be used for the sonification. The segment will be given to the composition algorithm which provides the mapping of color information to sound properties. The hue within a segment will be linked to the chords present in that defined scale. This is done by dividing the hue space into seven segments corresponding to the seven chords present in the scale, see Figure 31. The use of notes within a scale, instead of directly linking the hue space to pitch, is done to create pleasant and consonant sounds. Saturation will be linked to the loudness of the sound, meaning a pale color will have a quieter sound as opposed to a loud sound for deep colors. Lastly, value will be linked to the octave a segment is played in. With this in mind, if the average color of the overall painting is dark green, a Minor A will be selected as the scale of the composition. A bright blue, on the other hand, would set a Major E scale. If the average color of a segment is a dark deep green, the sixth chord of the scale will play loudly in a low octave. A bright pale blue, on the other hand, would play the third chord of the scale quietly in a high octave. An overview of the mapping from low-level visual features to sound properties can be found in Table 3.

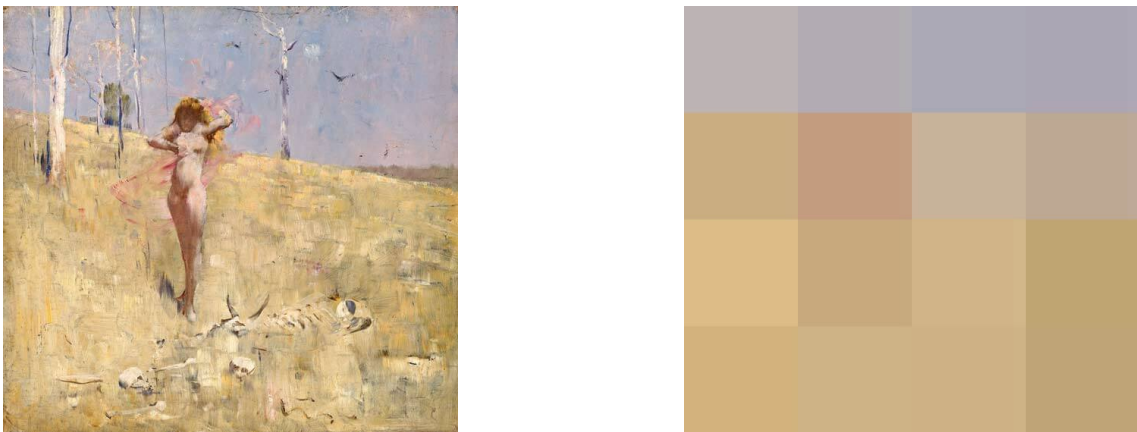


Figure 29: Painting by Arthur Streeton divided into segments

Visual feature	Audio property
Hue	Chord of scale
Saturation	Loudness
Value	Octave
Edge	Timbre

Table 3: Mapping of visual features to audio properties



Figure 30: Linking of hue and piano keys



Figure 31: Linking of hue to chords in a scale

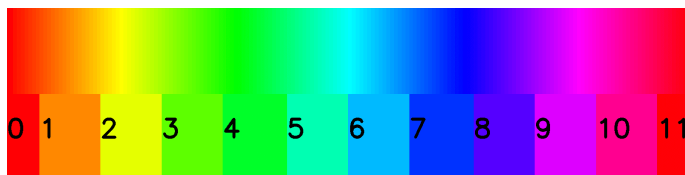


Figure 32: Scaling of the hue space to the scale-space

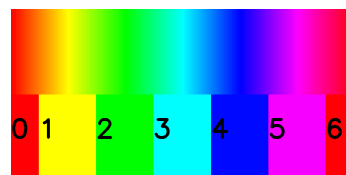


Figure 33: Scaling of the hue space to the chord space of a scale

5.1.3 Edges as timbre

To change the timbre of the sound based on the edges present within the segment waveshaping will be used. There are multiple ways a waveform can be shaped. By directly “drawing” a waveform, by the use of a transfer function to transform a predefined wave, or by defining an order of waveforms and morphing between them. The idea is to use the edge information present in a painting to influence the waveform produced by the means of waveshaping. This would lead to sharper sound for sharper edges.

See Figure 34, Figure 35, Figure 36 and Figure 37 for two examples of paintings and their edge-detected counterpart. Based on the two examples the advantages and disadvantages of the waveshaping techniques will be explained. One of the first visual findings is that paintings can inherently contain a lot of edges. This is probably because of the brush strokes within a painting.



Figure 34: Painting made by Bob Ross

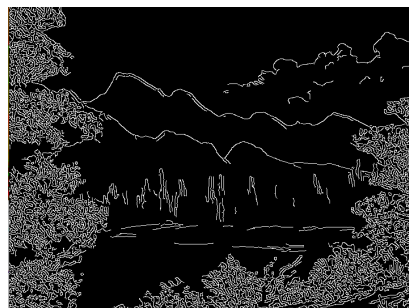


Figure 35: Edge detection on Figure 34

To be able to draw a waveform from edge information a line is extracted from each painting segment. This is done by a simple path-finding algorithm. This algorithm would traverse left

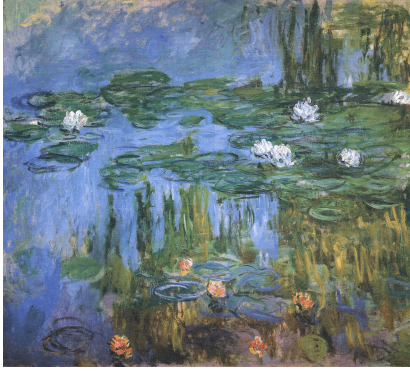


Figure 36: Painting made by Claude Monet

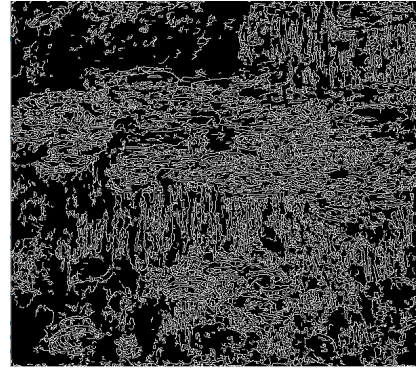


Figure 37: Edge detection on Figure 36

to right within a segment and look at a whole column of edge information, e.g. all y values at x position 0. The first step takes the closest edge point to the middle of the y -axes. Where in the next step, it would take an edge point closest to the previous point. If there was no edge information available, it would take the previous edge point value. A more in-depth explanation of the algorithm can be found in section 5.2. The result of this algorithm can be seen in Figure 38.

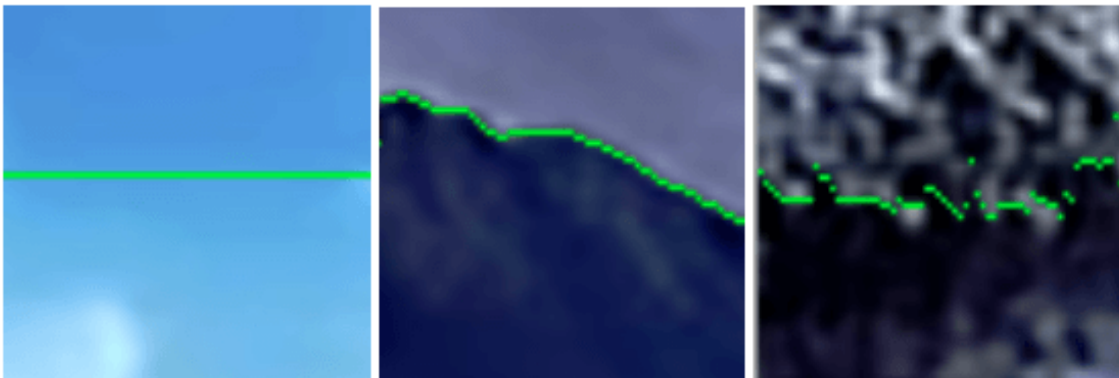


Figure 38: Results of the path-finding algorithm

When using this line data to directly draw a waveform a problem arises. When there is no edge detected the algorithm produces a flat line and a waveform has the shape of a flat line, it will not produce sound. The last example in Figure 38 produces a nearly flat line but this could be resolved by scaling the data. However, the fact that the line does not begin and end at the same Y values could also pose a problem. Waveforms that do not begin and end at the same point tend to produce harsh sounds, which might not always be desirable. Within this method, there is also the possibility that a line is not equally distributed. This can cause a positive or negative DC offset which can be harmful to speakers. Another problem shows when there is no clear line found in the edge information of the segment. This produces a noncontinuous line which could create a waveform close to noise, which is not musically pleasant.

The problem of a flat line not producing sound could be, at first thought, solved by waveshaping a sine wave. When there is no edge information a flat line should not shape the sine wave and therefore create a smooth sound. In such an implementation the line information would be used as a transfer function. However, if the transfer function is a flat line, the sine input would become a flat line, thereby not elevating the problems posed by directly drawing the waveform from the line information. One way to solve the above problems would be a more complicated path-finding algorithm that could make sure the line found in a segment would always produce a well-sounding waveform. The concept the edge information tries to solve is that a segment with more edges creates

a rougher sound. Therefore, another way to solve the problems would be to use a waveshaping technique having predefined waveforms and morph between them in a linear fashion. This offers a simple solution to the problem at hand. When there is no edge information a perfect sine will be produced, creating a smooth sound. If there is a lot of edge information present a Saw wave will be produced, creating a rough sound. Anything in between will produce the waveforms in between, creating a sound between smooth and rough. See Figure 39 and Figure 40.

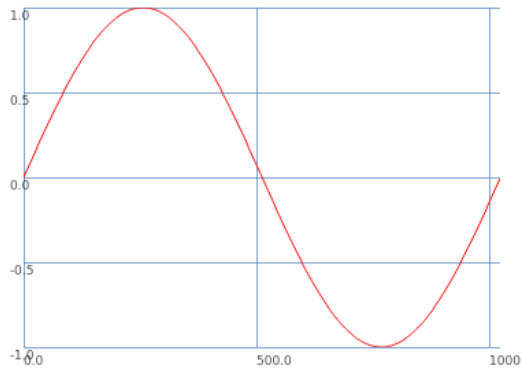


Figure 39: Sine wave



Figure 40: Saw wave

5.1.4 Histogram as timbre

To have a more unique sound per painting the histogram is used to create a waveform unique to the painting. To create a proper waveform the histogram, see Figure 42, is flipped horizontally and vertically and appended to the end of the existing line, see Figure 41. Using the histogram waveform as primary and only sound, there is no way for the edge to influence the timbre of the sound as there is no option to morph between waveforms anymore. Therefore a combination of waveshaping, based on edge and histogram information, has been implemented. It is based on the waveshaping method explained in Figure 5.1.2 by morphing between waveforms. The waveform produced by the histogram is added to the list of possible waveforms, see Figure 41. This is done with the idea that a sine wave sounds the smoothest because it only produces the fundamental frequency, and a saw wave would sound the roughest because it contains all the integer harmonics. With this assumption, the histogram wave will fit in between the other two waveforms based on roughness.

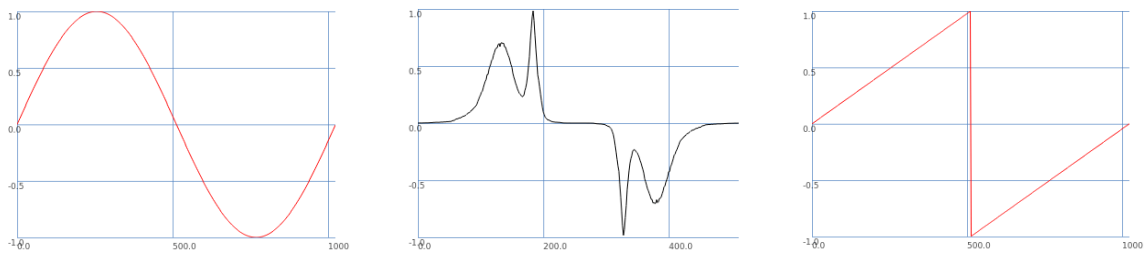


Figure 41: The waveforms used to change timbre based on the edge of a segment. Left: sine wave, Middle: waveform based on histogram (Figure 42), Right: saw wave

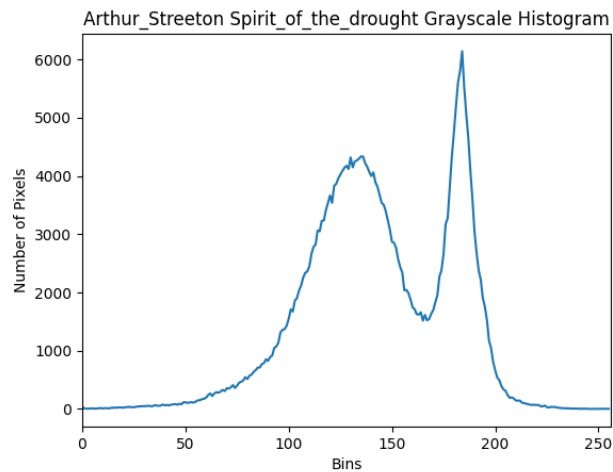


Figure 42: Grayscale histogram of a painting by Arthur Streeton

5.1.5 Edges as melodies

Although the line of the path-finding algorithm could not be used to shape a waveform, in a similar fashion as Yoshida et al. [43] it can be used to change the pitch of a sound. Therefore this line is used to create a melody that should convey the line information present in a segment. This is done by taking note samples of the produced line, where the lowest point represents the lowest possible note (0) and the highest point represents the highest possible note (6). The points in between are represented by the whole numbers between zero and six. This is in line with the seven notes available within an octave of a major or minor scale.

5.1.6 Panning based on location

To create a more dynamic musical piece, panning is used to create a sense of space. When the sonification of a segment is on the left side of a painting, the sound will also be panned to the left. The further away the segment is from the center of the painting, the more apparent the panning will become.

5.1.7 Navigation of segments: salience

For the sonification, the painting is divided into segments. One way to do the navigation is to let the segments play from left to right and top to bottom. However, this is not how people tend to look at paintings. Therefore the sonification is not done in a left to right top to bottom fashion, but rather in the order of saliency. The order is defined by how many salient pixels a segment has, meaning, the segment with the most salient pixels will sound first, and the segment with the least salient pixels last. To accomplish this the saliency map of a painting is created, see Figure 43. This map is then converted to a threshold map (Figure 44) to ease the counting process of salient pixels.

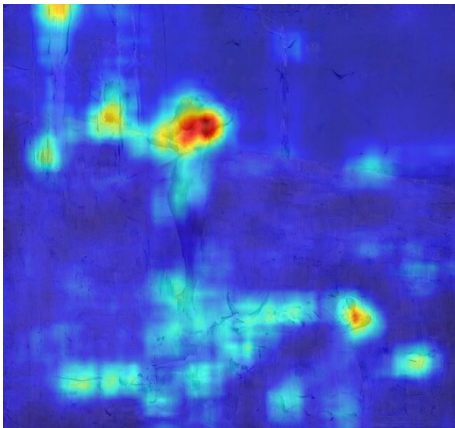


Figure 43: Saliency map based on a painting by Arthur Streeton



Figure 44: Threshold map based on Figure 43

5.2 Technical implementation low-level features: visual

To put all the above ideas into reality a framework has been created with the use of Python and Supercollider. Python is responsible for the image processing while Supercollider creates the sound. First, the visual side of the framework will be explained in more detail. Within the figures of this chapter, the gray and blue boxes represent a conversion or extraction of values. The gray boxes are only calculated one time, whereas the blue boxes are calculated for each individual segment. The purple and yellow boxes represent output values. The purple values are global values that stay the same throughout the sonification. The yellow values are unique for each segment.

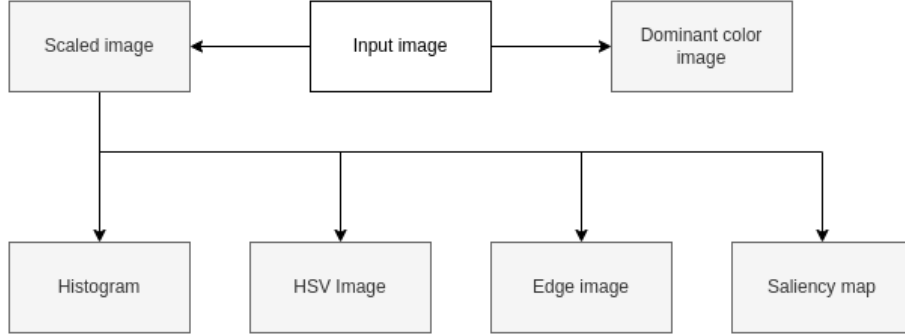


Figure 45: The extraction of the dominant color image, scaled image, histogram, HSV image, edge image, and saliency map

As a first step, the dominant color of the image is extracted and the image is scaled down, see Figure 45. The dominant color image is converted to HSV values, from these values the hue and value are extracted and scaled, see Figure 46. The hue of the dominant color will influence the root noted and is therefore scaled from the possible hue values of 0 to 179 to 0 to 11. This creates twelve possible outputs that correspond to the twelve possible notes within an octave. The equation used for scaling can be seen in Equation 1.

$$\frac{k_{min} + (x - j_{min}) * (k_{max} - k_{min})}{j_{max} - j_{min}} \quad (1)$$

In the above formula, the x represents the input value that is going to be scaled. j_{min} and j_{max} represent the range the input value can be in. E.g., x can be 0 at the minimum and 179 at maximum, then j_{min} will be 0 and j_{max} 179. k_{min} and k_{max} will define the output range of the scaled value. E.g., when the input value x is 90 and the input range is from 0 to 179, k_{min} can be set to 0 and k_{max} to 7 and the expected outcome will be 4. In the framework, the output values of the formula are rounded to create a less fine-grained scaling as floating points values would fall out of the scale the sonification needs to be in.

The value(V) of the HSV from the dominant color image is scaled from 0-255 to 0-1 using Equation 1. This is done as the value will influence the scale the sonification is played in. The two possible scales are minor (0) and major (1). A dominant color where the value of the HSV image is 127 or lower will create a minor scale sonification, whereas a value of 128 or higher will create a sonification in the major scale.

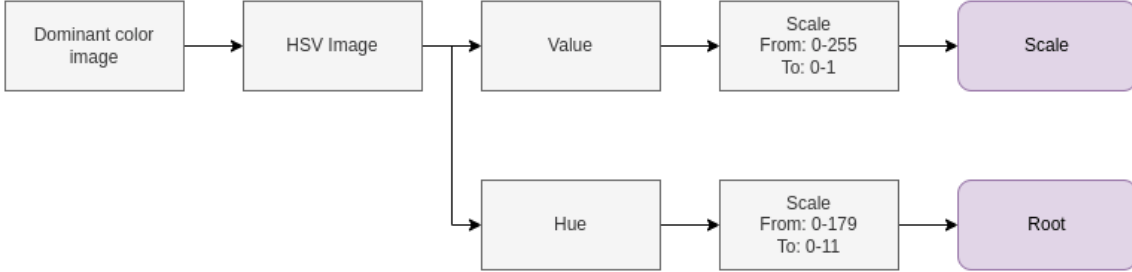


Figure 46: The use of the dominant color in the low-level visual framework

From the scaled-down image a histogram, the segmentation locations, and three new images are created, namely an HSV image, Saliency map, and an image containing all edges. The histogram is used to create a waveform. This waveform is later used to create a unique timbre for each painting, see section 5.1.4 for more high-level details. The segmentation locations are used to create a panning based on the location of the segment, see Figure 47. If the played segment is on the left of the center of the painting the panning of the sound will be left. The more a segment is to a certain side, the more the panning will be. The equation to calculate the panning can be seen in Equation 2.

$$\frac{p_{max}}{-2} + scale(step_{current} \bmod \sqrt{steps}, (0, \sqrt{steps - 1}), p_{min}, p_{max}) \quad (2)$$

The above formula uses the current step to calculate a panning. $step_{current}$ represents the current segment the framework is converting. $steps$ represents the total segments, or steps, the painting is divided in, e.g. 16. p_{max} represents the max possible panning value. This value is divided by -2 in the beginning of the formula to shift the output to a midpoint of 0. In the second part of the formula Equation 1 is used as a function to scale the step calculation to a meaningful panning value. p_{min} represents the lowest panning value, this value is likely 0 as the value gets shifted by the first part of the formula. As panning only occurs from left to right in a stereo audio output, $step_{current} \bmod steps$ is needed to only calculate the panning based on the x position. E.g., if there is a 4*4 grid the values should range from 1 to 4 and not 1 to 16.

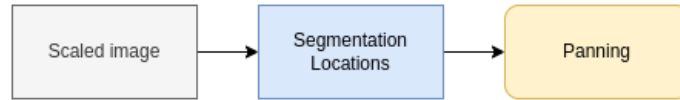


Figure 47: The creation of panning by the low-level visual framework

The edge image is used for two purposes. One of the functions is to calculate the edginess of a segment as a percentage, meaning the number of pixels containing an edge will be counted and divided by the total amount of pixels present in a segment, see Equation 3.

$$\frac{p_{255}}{p_{total}} \quad (3)$$

In the above formula, p_{255} represents the pixels that have a value of 255, meaning they are part of an edge. p_{total} is the total amount of pixels present in a current segment.

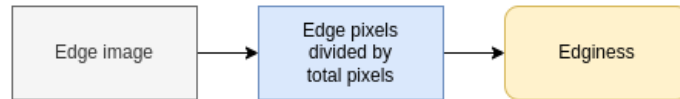


Figure 48: Calculation of edginess by the low-level visual framework

The other function is to create a line from the edges present in a painting. This line will be used by the audio generation side of the framework to create a melody, see Figure 49. The algorithm used to create a line from the edges in a segment can be seen in Listing 1.



Figure 49: Creation of the melody points by the low-level visual framework

```

1  start_position = int(math.floor(len(edge_segment) / 2))
2  line = []
3  for i, col in enumerate(edge_segment.T):
4      edge_positions = np.where(col == 255)[0]
5      if len(edge_positions) > 0 and len(line) > 0:
6          close = min(edge_positions, key=lambda pos: abs(pos - line[-1]))
7          line.append(close)
8      elif len(edge_positions) > 0 and len(line) == 0:
9          close = min(edge_positions, key=lambda pos: abs(pos -
10         start_position))
10         line.append(close)
11     elif len(line) > 0:
12         line.append(line[-1])
13     else:
14         line.append(start_position)
15
16     inverted_line = [len(edge_segment) - p for p in line]
17     scaled_inverted_line = []
18     for point in inverted_line:
19         scaled = scale_between_range(point, (0, len(edge_segment)), (0, 11)
20     )
21     scaled_inverted_line.append(scaled)
  
```

Listing 1: Line Algorithm

To find a line in a segment, this algorithm starts at the middle point of the y axis and the 0 point of the x-axis. As a second step, the algorithm will go through each column, finding an edge pixel closest to its current y position. This has the effect that the algorithm can produce an interrupted line, meaning neighboring line pixels are not always adjacent (Figure 38), this is however no problem for the use case of this line as a melody. Because of the data structure the CV2 library uses for images, the $x=0, y=0$ point of an image is in the top left and the y axis increases in steps the lower

you go. Therefore, the created line needs to be inverted to make it match the intuitive structure of a high point playing a high note. This is the function of the last part of the algorithm.

From the saliency map a threshold image is created, see Figure 50. From this threshold image, the percentage of salient pixels of each segment is calculated. This percentage is used to define the order in which the segments are played. The segment with the highest percentage is played first and the segment with the lowest percentage is played last. The equation can be seen in Equation 4

$$\forall s \frac{s \cap p_{255}}{(s \cap p_0) \cup s(\cap p_{255})} * 100 \quad (4)$$

In this formula, s is a set containing sets of all the pixels of each segment. p_{255} contains all pixels having a value of 255, and are therefore salient pixels. p_0 are the pixels having the value 0, therefore containing all the non-salient pixels. The formula divides the salient pixels by the total amount of pixels in a segment. The output of this division is multiplied by 100 to create an easily readable percentage. The output of the formula is a set of saliency percentages of each segment that can be used to create a navigation order.

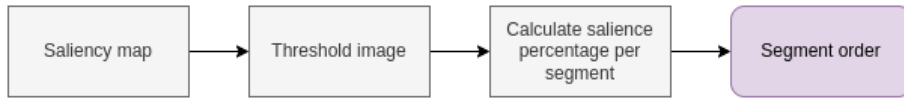


Figure 50: Creation of the navigation order by the low-level visual framework

From the HSV image the hue, value, and saturation are extracted, see Figure 51. The hue value is scaled, using the formula of Equation 1, from the possible range of 0 to 179 to a range of 0 to 6. The range of 0 to six corresponds to possible chords available within an octave. Saturation is scaled from the possible range of 0 to 255 to a range of 1 to 100 to influence the volume of the segment in the sonification. Value is scaled from the possible range of 0 to 255 to a range of 2 to 5 to influence the octave a segment is played in. This is done so that the sound of the sonification does not go too high or too low to the point where it becomes inaudible or unpleasant.

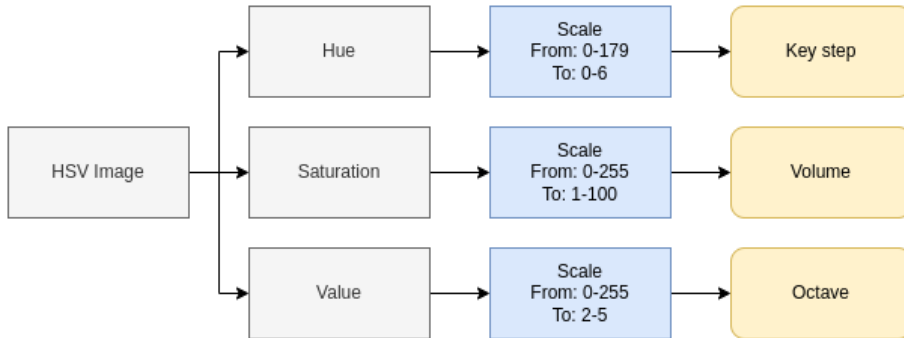


Figure 51: Extraction and scaling of the HSV values by the low-level visual framework

5.3 Technical implementation low-level features: audio

The visual part of the framework scans through the image and creates parameters that can be used by the audio side of the framework to create sound. The Sound Engine is created in SuperCollider and reads the parameters created by the visual side of the framework from a generated text file. At startup, new parameters and synth definitions are created and the parameters from the text file get loaded into memory, see Figure 52. At first a sine wave and saw wave are created. After that the waveform created for the histogram is loaded. The three waveforms are combined in a set so they can be used by the wavetable synth. After the waveforms are ready the synth definitions for the wavetable synth and the reverb are created. The definition of the wavetable synth can be seen in Listing 3. When the synth definitions are ready the parameters generated by the visual side of the framework are loaded into memory. Some parameters can be used as-is, whereas some parameters need to be transformed to be usable within the Sound Engine. The melody points are saved as one line within the generated text document. This would result in a lot of melody notes, which is undesirable. Therefore the number of melody notes can be given. The algorithm shown in Listing 2 takes the desired amount of samples from the melody points in the text file in an evenly distributed manner to create a simpler melody. Furthermore, the algorithm replaces a melody note with a rest note when the consecutive note is the same as the previous note. This is done to mitigate the possibility of an uninteresting melody containing a lot of similar consecutive notes. When all the variables are ready and loaded, they are given to the Pbinds for the bass, chords, and melody. Pbind is a function of SuperCollider that is used to play notes based on given parameters. As an example, the Pbind created for the framework to play chords can be seen in Listing 4. An overview showing the flow of the audio part of the low-level side of the framework can be found in Figure 52.

```
1 melody_note_count = 0;
2   steps.do({
3     arg s;
4     var tempDurArr = Array.newClear(melody_notes_amount_local - 1);
5     melody_notes_amount_local.do({
6       arg i;
7       var randDur = exprand((durations[s] / melody_notes_amount_local) - 0.1,
8         (durations[s] / melody_notes_amount_local) + 0.05);
9       var lastDur = 0;
10      var melodyNote;
11      var newMelodyNote;
12      melody_notes_durations[melody_note_count] = durations[s] /
13      melody_notes_amount_local;
14      melodyNote = lines[s][(i * (lines[s].size / melody_notes_amount_local))
15      .trunc];
16      if(melodyNote == lastMelodyNote) {newMelodyNote = \rest;} {
17      newMelodyNote = melodyNote};
18      melody_notes[melody_note_count] = newMelodyNote;
19      lastMelodyNote = melodyNote;
20
21      melody_octaves[melody_note_count] = octaves[s] + 1;
22      melody_amps[melody_note_count] = amps[s] + 0.08;
23      melody_pans[melody_note_count] = pans[s] + rrand(-0.1, 0.1);
24
25      melody_note_count = melody_note_count + 1;
26    });
27  });
```

Listing 2: Line to melody algorithm


```

1 SynthDef.new(\vosc, {
2   //Parameters
3   arg buf=0, numBufs=1, bufPos=0,
4   freq=440, atk=0, dec=1, sus=0, rel=0.2,
5   amp=0.2, gate=1, pan=0,
6   out=0, fx=0, fxsend=(-25);
7
8   //Create inline variables
9   var sig, detuneSig, env;
10
11  //Calculate the sound
12  bufPos = buf + bufPos.min(numBufs - 1).max(0);
13  env = Env.adsr(atk, dec, sus, rel);
14  detuneSig = LFNoise1.kr(0.2!2).bipolar(0.2).midiratio;
15  sig = VOsc.ar(bufPos, freq * detuneSig);
16  sig = Splay.ar(sig, center:pan);
17  sig = LeakDC.ar(sig);
18  sig = sig * EnvGen.kr(env, gate, doneAction: Done.freeSelf);
19  sig[1] = DelayN.ar(sig[1], 0.016, 0.016);
20  sig = LPF.ar(sig, freq: 16000, mul: 1.0, add: 0.0);
21  Out.ar(0, sig * amp);
22  Out.ar(fx, sig * fxsend.dbamp);
23 }).add;

```

Listing 3: Wavetable synth definition

```

1 Pbind(
2   //Pbind parameters to play notes
3   \instrument, \vosc,
4   \degree, Pseq(chords),
5   \root, root,
6   \octave, Pseq(octaves),
7   \dur, Pseq(durations),
8
9   //Parameters given directly to the synth definition
10  \amp, Pseq(amps),
11  \atk, Pseq(durations * 0.5),
12  \dec, Pseq(durations - releases),
13  \sus, 0,
14  \rel, Pseq(releases),
15  \buf, buf[0].bufnum,
16  \numBufs, buf.size,
17  \bufPos, Pseq(edginess),
18  \pan, Pseq(pans),
19  \fx, ~vbus,
20  \fxsend, -20,
21 ).play;

```

Listing 4: Pbind used to play the chords

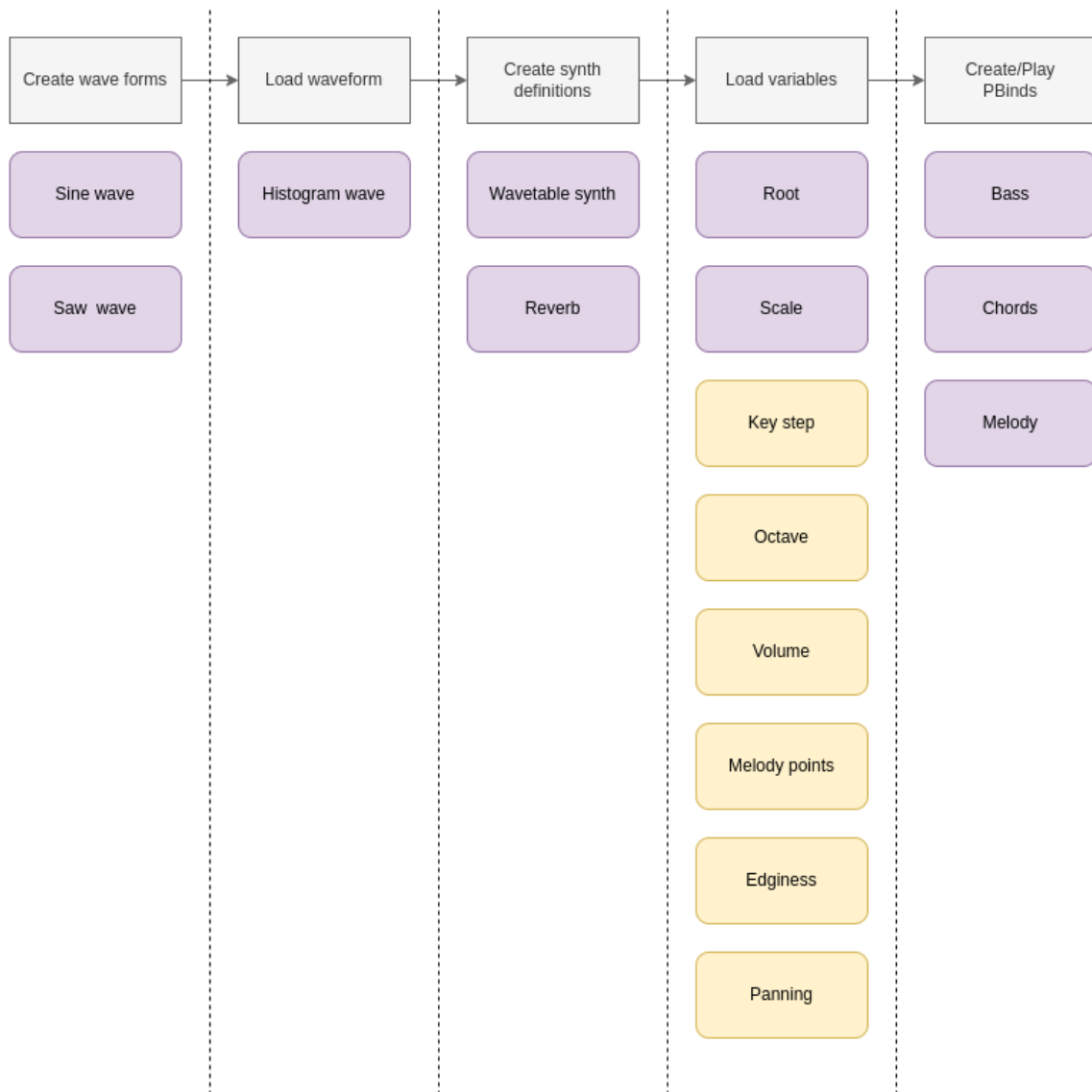


Figure 52: The flow of the audio part of the frameworks low-level side

5.4 High-level feature sonification

To answer the question “How can a sonification pipeline be created to incorporate high-level features extracted from paintings?” inspiration has been taken from Rector et al.[32]. Rector et al. use the sound of objects in the painting to convert these high-level features into sound. Instead of looking at individual objects, this framework tries to convey the whole scene of the painting. To accomplish this the Places[44] model has been used to extract the scene information of a painting. This information is linked to a corresponding sound for that particular scene, e.g. when the painting inhabited the scene of a market, the sounds heard at a market would be added as a background layer in the sonification.

5.4.1 Technical implementation

To create a high-level features implementation some additions to the framework are made. On the visual side, Places is been used to extract the scene label from the painting. To match up the scene label with a sound, first scene labels need to be created for an audio dataset. To do this in an automatic manner Soundnet[1] has been utilized to label large audio datasets. The datasets were ESC-50[27], FSD50K[11], TUT acoustic scenes 2016[23], and TUT acoustic scenes 2017[24] because of their scenic nature. In the matching process, when more than one match has been found, a random sound will be chosen from all the matches. The output is a path to the audio file that will be added to the generated text file containing the other parameters for the sonification.

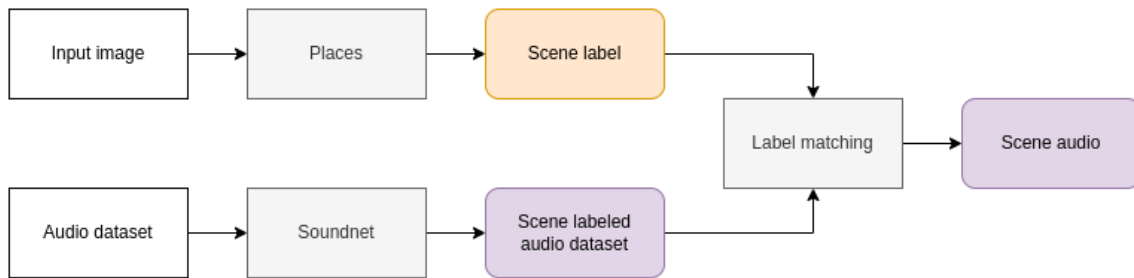


Figure 53: Extraction of the scene label from a painting with the Places model and an audio dataset with the Soundnet model

In the Sound Engine, a synth definition is added to be able to play the scene audio files. There is a version that can play mono audio files and one that can play stereo audio files. As an example, the stereo version is shown in Listing 5. Instead of using Pbind to play the synth definition, for playing the audio file an instance of the synth is directly created, see Listing 6. This is because the audio file should play in its original pitch and speed and is, therefore, a synth that should not play different notes. Therefore there is no need to use an intermediate function such as Pbind.

```

1 SynthDef(\audioFileStereo, {
2   arg out, fx=0, fxsend=(-25), amp=1, bufnum, sustainTime,
3   atk=0, dec=1, sus=0, rel=0,
4   gate=1, rate=1, t_trig=1, start=0, loop=1;
5   var sig = PlayBuf.ar(2, bufnum, BufRateScale.kr(bufnum) * rate, t_trig, start
6   , loop);
7   var gateEnv = EnvGen.kr(Env([1, 1, 0], [sustainTime, 0]));
8   var env = EnvGen.kr(Env.adsr(atk, dec, sus, rel, 1), gate * gateEnv,
9   doneAction: Done.freeSelf);
10  sig = CompanderD.ar(sig, sig, thresh: 0.4, slopeBelow: 0.5, slopeAbove: 0.1,
11  clampTime: 0.01, relaxTime: 0.01);
12  sig = FreeVerb2.ar(sig, sig, mix: 0.5, room: 0.2, damp: 0.5);
13  sig[0] = DelayN.ar(sig[0], 0.024, 0.024);
14  sig = LPF.ar(sig, freq: 4000);
15  sig = HPF.ar(sig, freq: 220);
16  Out.ar(out, sig * env * amp);
17 }) .add;

```

Listing 5: Stereo Audio synth definition

```

1 scene_buffer = Buffer.read(s, scene_path, action: {wait_for_scene_load = false
2   ;});
3 scene_buffer = scene_buffer.normalize;
4 scene_args = [
5   \bufnum, scene_buffer.bufnum,
6   \atk, 4,
7   \dec, 0,
8   \sus, 1,
9   \rel, 4,
10  \sustainTime, durations.sum,
11  \amp, max(amps.minItem() + 0.12, 0.0006)];
12 Synth.new(\audioFileStereo, scene_args);

```

Listing 6: Code to play scene audio

5.5 Description of the sound

When listening to the model one can think of the genre ambient. The resulting musical pieces do not contain any drums and portray a certain type of calmness and waviness. The sound can feel like it is in the distance and floating in the air. The calmness can be attributed to the slow and smooth attack and the decay of the sound and the slow progression of notes. Because of the digital nature of the sound, the timbre has a distinctive digital flavor. Although the sound is digital it still can portray a certain kind of warmth.

5.6 Intermediate results

After looking at the results of the first model a few problems became clear. The labeling of the scene audio datasets by Soundnet was not always accurate. The input datasets used were ESC-50[27], FSD50K[11], TUT acoustic scenes 2016[23], and TUT acoustic scenes 2017[24]. These datasets are chosen as they contain a lot of scene sounds. Although these datasets contain a lot of scene sounds, they, unfortunately, did not span all scene categories from the painting dataset. As the performance of scene detection on audio was not the focus of this research a manual scene audio dataset was created. This was done by means of filtering the noise from the automatically created dataset by Soundnet and filling the missing categories manually. Another issue that arose was the current segmentation method. With the square segmentation design, the border of different objects was not taken into account. Therefore, a segment could contain half a head of a person in a painting with a bit of sky. This type of segmentation felt unnatural and it was, therefore, decided the segmentation method would change in the second model. However, the most prominent problem is diversity. Most of the sonifications tend to sound very familiar, making it hard to match paintings with their sonifications without the help of the scene sound. To combat this problem, a new synthesis method will be tried in the next iteration of the model.

6 Model 2: object segmentation and FM synthesis

The second model is based on the first model with a couple of changes in the visual processing and the audio generation. The first thing that stood out in the first model was the square segmentation method. Dividing the painting into squares felt unnatural, therefore a different method of segmentation was explored. On the audio side, it was noticeable that the diversity of the created sounds was low within and between paintings.

6.1 Object segmentation from high-level features

Instead of diving a painting into squares an approach similar to the segmentation by Rector et al. [32] was implemented. Rector et al. made a manual divisions of a painting into natural segments e.g. the sky or a person would form a segment separately. To implement such an approach automatically, Detectron2[40] was used to divide a painting into natural segments, see Figure 54. The category classification of each segment was ignored and only the dominant color of each segment was used in the sonification, see Figure 55. This choice was made because the dominant colors are more representative than the average of the colors present within a segment. Besides the dominant color of a segment, the size was also taken into account. To convey the size of a certain segment the percentage of pixels within the segment resulted in a change of duration of a specific segment in the sonification, see Equation 5. In the equation, $p_{segments}$ represents the number of pixels of a segment, and p_{total} represents the total amount of pixels within the painting. This division will represent the percentage of space a segment takes up within a painting. This percentage is scaled to a suitable note duration. A small segment will therefore sound 2 seconds, while a large segment will sound 8 seconds. The result showed more of a more natural segmentation of the painting but, had little influence on the sound produced aside from the fact that sonification now could have varying durations. The change from segmentation method also had an influence on the panning, as the current equation used to calculate the panning assumed a square pattern-like structure of segments.

$$scale\left(\frac{p_{segment}}{p_{total}} * 100, (0, 100), (2, 8)\right) \quad (5)$$



Figure 54: Natural segmentation by Detectron 2 of La Piazzetta by Corot Venice



Figure 55: Dominant colors of segments in La Piazzetta by Corot Venice

6.1.1 New panning

The new panning method looks at the position of the segment relative to the middle of the painting. If more pixels of the segments are on the left the more the panning will be to the left. If the segment is entirely on the left of the painting the panning to the left will be at its maximum. The higher the percentage of the segments pixels are on the right the more the panning will be to the right. To calculate the panning Equation 6 is used. In this equation, p_{total} represents the total amount of pixels present in the segment. p_{left} is the amount of pixels from the segment that are on the left of the painting relative to the middle. p_{right} represents the number of pixels that are on the right.

$$- 1 * \frac{p_{left}}{p_{total}} + 1 * \frac{p_{right}}{p_{total}} \quad (6)$$

6.1.2 Segments as melody

The change from square segments to natural segments also has an impact on melody creation. In the first model, the melody is created from a line drawn within a square, however because of the undefined shape of the new segments this option is no longer valid. Segments inherently follow the edges of objects, therefore drawing a line within such a segment seems counterintuitive. Thus, instead of taking a line to represent the melody of a segment, all segments of a painting are combined to form a melody pattern, based on the colors of the segment. This pattern will then be shuffled each time a segment is played. This with the idea to convey all the colors present within the painting.

6.2 FM synthesis

In the first model, Wavetable synthesis was used to provide a way to convey the roughness of a painting. This was done by providing three waveforms, namely a sine wave, a waveform extracted from the painting's histogram, and a saw waveform. The number of edge pixels within a segment was used to linearly interpolate between the before mentioned waveforms to convey roughness. One downside of this implementation was that most results sounded alike. Therefore another synthesis method, namely FM synthesis, was explored. FM Synthesis works by having two waveforms that can influence each other. One of the waveforms is called the carrier and the other waveform is called the modulator. In the most basic form, the two waveforms are sine waves, where the modulator wave modulates the carrier wave, see Figure 56. This opens up a new array of parameters that can be changed based on color or other visual information. It also creates the possibility to have the carrier or modulator waveform be represented by the wavetable used in the previous model to create even more parameters. In the second model FM synthesis is implemented by having the wavetable from the first model as the carrier and a sine wave as the modulator, see Listing 7. However, in contrast to the first model, the wavetable does not interpolate between the waveforms anymore, as this made the sound too complex. Instead, it changes to one of its three waveforms in a threshold manner. E.g. when the percentage of edge pixels in a certain segment exceeds a threshold a certain wave is used. The thresholds chosen for this model are 0-0.33 (first waveform), 0.33-0.66 (second waveform), and 0.66-0.99 (third waveform).

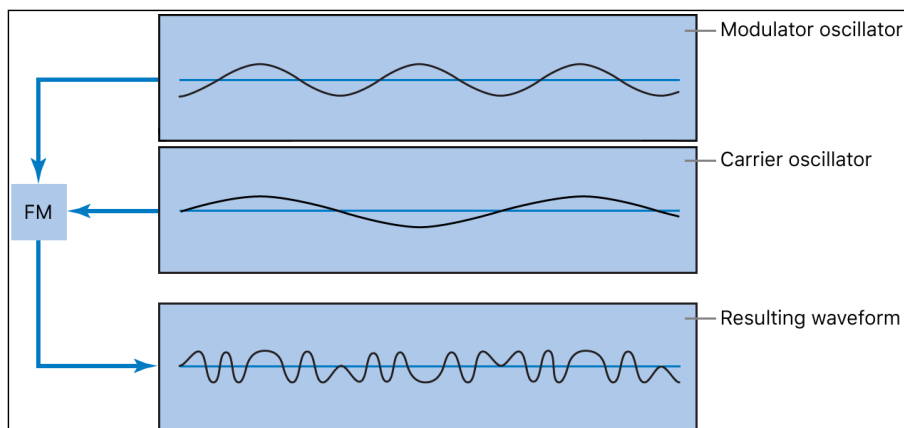


Figure 56: Example of FM-synthesis with two sine waves

```

1 SynthDef(\fm, {
2   arg buf=0, numBufs=1, bufPos=0,
3   freq=500, mRatio=1, cRatio=1,
4   index=2, iScale=2, cAtk=4, cRel=(-4),
5   amp=0.2, atk=0.01, sus=0, dec=1, rel=0.2,
6   pan=0, out=0, fx=0, fxsend=(-25);
7   var car, mod, env, iEnv, detuneSig;
8   bufPos = buf + bufPos.min(numBufs - 1).max(0);
9   detuneSig = LFNoise1.kr(0.2!2).bipolar(0.2).midiratio;
10  //index of modulation
11  iEnv = EnvGen.kr(
12    Env(
13      [index, index*iScale, index],
14      [atk, dec, rel],
15      [cAtk, 1, cRel],
16    ),
17    doneAction:Done.freeSelf
18  );
19  //amplitude envelope
20  env = EnvGen.kr(Env.adsr(atk, dec, sus, rel), doneAction:Done.freeSelf);
21  //modulator/carrier
22  mod = SinOsc.ar(freq * mRatio, mul:freq * mRatio * iEnv);
23  car = Osc.ar(bufPos, freq * detuneSig * cRatio + mod) * env * amp;
24  car = LPF.ar(car, freq: 2000, mul: 1.0, add: 0.0);
25  car = Pan2.ar(car, pan);
26  Out.ar(out, car);
27  Out.ar(fx, car * fxsend.dbamp);
28  }).add;

```

Listing 7: FM synth definition

6.3 Description of the sound

While in many ways this model shares a lot with the first model, there are some differences in the sound. Because the played chords and notes do not have a fixed duration now, the results can inhibit tempo changes, thereby making some sonifications sound less calm, but more dynamic. Furthermore, because of the change of synthesis method, this model also tends to sound brighter. While the sound can still be considered soft overall it is a bit harsher than the first model.

6.4 Intermediate results

Looking at the result of model 2, the following stood out: the new segmentation method that was implemented was not always perfect, see Figure 57. However, as the focus of this research was not to improve the segmentation of paintings, the results were considered good for the aim of the framework.



Figure 57: Example of a noisy segmentation. On the left is the segmentation by Detectron2 on Paul Delvaux - The Viaducto. On the right are the dominant colors used for the sonification

Apart from that, the problem with diversity still existed. Trying Fm-synthesis did not open as many useful parameters as originally hoped for. When linking output parameters from the visual side of the framework to new parameters available on the newly created FM-synth definition the sound quickly became unpleasant. This problem could be caused by the implementation of the FM-synthesis and might not be directly related to its concept. The only new linked parameter was from the edginess of the visual side to the index of the FM-synth on the audio side. Therefore, the problems with diversity still arose in the results of the second model.

7 Model 3: Instruments

Because the second model still exhibited problems with diversity a third model has been designed. This model takes inspiration from the method used by Cho et al. [5]. Cho et al. took the approach of linking certain colors to certain instruments playing a certain chord to create a noticeable difference between colors. In this model the same approach has been taken where colors are not only linked to certain chords within a scale, but every color has its own unique instrument. To be able to play instruments, first, the sound samples of the instruments are needed to be loaded into buffers within the Sound Engine. To create the sound of the instruments a synth definition has been created that can play the instrument samples from the loaded buffers, see Listing 8. In the Pbind, the function responsible for playing notes based on given parameters, the linking of the hue to an instrument is created, see Listing 9.

```
1 SynthDef(\instruments, {
2   arg out, freq=60.midicps, amp=1, buf, sustainTime,
3   atk=0, dec=1, sus=0, rel=0,
4   pan=0, gate=1, t_trig=1, start=0, loop=1;
5   var sig = PlayBuf.ar(2, buf, (freq / 60.midicps) * BufRateScale.kr(buf),
6   t_trig, start, loop);
7   var gateEnv = EnvGen.kr(Env([1, 1, 0], [sustainTime, 0]));
8   var env = EnvGen.kr(Env.adsr(atk, dec, sus, rel), gate * gateEnv,
9   doneAction: Done.freeSelf);
10  sig[0] = DelayN.ar(sig[0], 0.024, 0.024);
11  sig = Splay.ar(sig, center:pan);
12  Out.ar(out, sig * env * (amp + 0.4));
13 }
```

Listing 8: Instrument synth definition

```
1 Pbind(
2   \instrument, \instruments,
3   \degree, Pseq(chords),
4   \root, root,
5   \octave, Pseq(octaves),
6   \dur, Pseq(durations),
7   \amp, Pseq(amps + 1),
8   \atk, Pseq(durations * 0.8),
9   \dec, Pseq(durations + 3),
10  \sus, 0,
11  \sustainTime, Pseq(durations),
12  \rel, Pseq(releases),
13
14  //Create a set by selecting the right instrument for a certain hue
15  \buf, Pseq(hues.collect({arg i; instruments[i]})),
16
17  \pan, Pseq(pans),
18  \fx, ~vbus,
19  \fxsend, -10,
20 ).play;
```

Listing 9: Instrument pbind

7.1 Description of the sound

The overall timbre of this model differs from the other two models as the sound is not synthesized but is created by using samples of actual instruments. Although this might lead one to think that the sound would have a less digital feeling, this is however not the case. The sound of this third model still has a very digital feel. Also, compared to the other models it feels less warm. This can be caused by the fact that the instruments are played digitally and not by actual humans, thereby losing a human touch. The sound also feels more direct and less grand than the first and second models, thereby making it feel more in the foreground. This can be due to the fact that the sound is simpler as the timbre of the notes of a single segment is produced by one instrument. Furthermore, the instrument playing the notes changes based on the color of a segment. This creates a feeling of separation, where each segment represents a single part. This is in contrast to the first and second models, which sound more like a whole.

7.2 Intermediate results

Although the diversity in the sound increased by using a different instrument for a specific hue, the composition of the chords was still the same. From the result, it was still hard to differentiate one painting from another. Another problem that arose from only using instruments was that the sound started to sound quite flat and less dynamic than previous models. To solve the flatness of the sound the next model will be a combination of the second and third models. In a last attempt to solve the diversity problem a couple of changes will be done on the visual processing side of the framework.

8 Model 4: Instruments accompanied by FM synthesis

The fourth model is a combination of the second model and the third model in order to get the best of both models. Furthermore, some changes have been done on the visual side. Inner scaling is introduced to create more diversity in the notes played within the sonification. Also, objects are now used to create a sense of chaos by linking the object amount present in a painting to the duration of notes. Thereby a painting with more objects will create a more chaotic sonification.

8.1 Inner scaling

In the previous models, values were scaled based on the minimum and maximum of the input value. E.g. hue has a possible range of 0 to 127 in the CV2 library. Therefore, this hue value is scaled down from 0 to 127 to a range of 0 to 7 to accommodate the possible chords within a scale. This is quite a big reduction in resolution. Therefore, if a painting contains a lot of colors that are close to each other in e.g. the hue, this big reduction will give a high chance that these colors will play the same notes. To enhance the diversity Inner Scaling is created. Inner Scaling is a way to minimize the resolution reduction by limiting the possible input values. E.g. by only taking the hue values present in the painting, the possible input values will be reduced. A downside of this is that the scaling will be dynamic and no color is always linked to a certain sound. Meaning, when scaling with Inner Scaling the linking of e.g. hue to chords can differ per painting, making it hard to extract specific colors across paintings, even with training.

8.2 Objects as an influence in note duration

In an effort to create more diversity between the sonifications a high-level feature parameter will be used to influence the duration of the notes, see Equation 7. By using the number of objects in paintings a better distinction can be made between calm and chaotic paintings. This is done with the assumption that a calm painting will contain lesser objects than a chaotic painting and long slow notes are considered calmer than quick short notes.

$$scale\left(\frac{p_{segment}}{p_{total}} * 100, (0, 100), \left(\max\left(1, 4 - \left(\frac{objects}{4}\right)\right), \max(4, 16 - objects)\right)\right) \quad (7)$$

Equation 7 is based on the same principle used in Equation 5 with the addition of a dynamic output range. When more objects are in the painting the minimum and the maximum output will be lower, when there are fewer objects the output will be higher, resulting in a longer duration.

8.3 Description of the sound

The fourth model is a combination of the second and the third model. Therefore the sound of the fourth model sounds warmer and softer than the third model. Also, this model sounds grander and more spacious, as it gets that property from the second model. Because of the combination between the second and third models, the sound is more like a whole, but one can still discern the feeling of parts of the third model, making it more dynamic.

8.4 Intermediate results

With the new changes in place, the last model tends to exhibit a bit more diversity than earlier models. However, this diversity shows mostly within the sonification itself and not between paintings. Another problem is that Inner Scaling while giving the sonifications a more dynamic composition, it does not handle paintings with a small amount of contrast well. E.g. when a painting consists of only bright colors, the darkest color will still be played in the lowest octave because of the dynamic scaling. This can produce an undesirable result. The note duration based on the number of objects in the paintings tends to work pretty well, however, it is not perfect. The assumption that a painting is more chaotic when it contains more objects does not always hold. Overall the musical pieces tend to be more dynamic, containing less of the same notes on chords. However, it still seems to be difficult to differentiate between sonifications and match a sonification with its painting. However, the performance until this point has only been evaluated by three participants and the main researcher. To evaluate the performance further, interviews with experts will be conducted. The purpose of these interviews is to find new solutions to current problems, to find problems that have not yet been identified, and to find inspiration to extend and improve the current framework.

9 Evaluation by experts

The evaluation consisted of interviewing experts on the output of the created models. The main goal of the interviews was to get expert feedback on the performance of the last model, and what improvements are possible in their opinions. However, as the time of experts is assumed to be limited the decision has been made to only evaluate the 4Th model. This model was considered the most advanced and promising by three participants and the main researcher. To guide the interview a dataset and a couple of questions were prepared.

9.1 Dataset

To create a sensible dataset for this research the dataset will consist of a filtered version of the [Painter by Numbers](#) dataset. It was decided to filter the dataset on Impressionism because of the close relation to natural images. The assumption is that the performance of scene detection will be the best with paintings close to natural images. After filtering the dataset consisted of 8279 paintings in 29 genres, see Appendix A Figure 74. To further downsize the dataset the following genres have been filtered out: caricature, illustration, nude painting (nu), panorama, portrait, poster, self-portrait, sketch, study, still life, symbolic painting, vanitas. The reason for the omission of the above genres is because of their lack of color or definable scene. After filtering the dataset contains 6514 paintings, see Appendix A Figure 75. The next step is to see how scene detection performs on the selected paintings, however, 6514 paintings would take a long time to evaluate manually. Therefore, a maximum of 10 paintings per genre is randomly sampled, leaving 143 paintings, see Appendix A Figure 76. After evaluating the dataset consisting of 143 paintings on top-1 scene prediction correctness 44 paintings remain, see Appendix A Figure 77. After looking at the 44 paintings it became clear that a lot of paintings looked similar, therefore the decision has been made to make a smaller selection from the current 44 paintings and add some paintings from outside the Painter by Numbers dataset. The result was a dataset containing 8 paintings where most differ in color and scene. All paintings can be seen in Appendix B.

9.2 Interview questions

For the interviews, a semi-open interview format is chosen to keep a similar flow between experts. This makes sure analyzing and comparing the interviews is possible, but also gives freedom to focus or divert into other topics if the interviewer or expert finds it important to do so. Therefore, a couple of questions have been predefined.

The first two questions regarding the sonification topic are "How would you turn this painting into sound?" and "Is there a type of model/technique would you use?". These questions are to provide any interesting insights into the way painting sonification can be done before participants are influenced by the sonification method of this research.

After those questions are answered, participants are asked to listen to the sonification of the eight paintings in the evaluation dataset and to answer the following questions. "What mental image do you get from listening to this song?", "What part of the song influenced your mental picture the most?" and "Would you describe the song you heard as pleasant?". The purpose of the first two questions is to extract the mental image experts got while listening and because of what aspect they got this image. The answer to these questions could be used to see how they would overlap with the content of the painting. The last question was asked to see if the participant would describe the sonification as pleasant.

After listening to the paintings the experts were shown each of the eight paintings and asked the following question: "What sound/song do you imagine when looking at the painting?" With the answer to this question a comparison can be made between what the participant imaged and what it would create as a musical piece. Also, inspiration for further improvement of the framework can be taken from the answers to this question.

To test the descriptiveness of the sonifications experts were presented with two paintings and one sonification. The sonification originated from one of the two paintings shown. Experts were asked the following questions: "Could you choose one of the two paintings you find best fitting to the musical piece you heard?", "Could you explain your choice?", and "What aspects of the song or the painting stood out or influences your choice the most?"

After testing the descriptiveness, the experts would be looking at a painting and hearing the corresponding sonification simultaneously. After each painting, the following questions were asked: "Do you think the song is descriptive of the painting?", "Could anything be added or changed?", and "What are most the most descriptive features of the song to you?" The purpose of the questions was to see if the expert found the sonification descriptive of the painting and if or how the expert would improve or change the sonification.

9.3 Results per painting

The results of the interviews will be discussed per painting and per question in this section. The sonification experts evaluated has been added as a link under the corresponding figure. Because of the semi-open interview format, the researcher could ask follow-up questions or questions that are not in the predefined list of questions. Also, the expert could freely provide information not related to the question. This information will be discussed in section 9.5

9.3.1 Charge of the scots greys at waterloo



Figure 58: Charge of the scots greys at waterloo

<https://drive.google.com/file/d/1iMBV5mPiJ17YLgzEbcIvuP0r7d4sAGsL/view?usp=sharing>

While listening to a sonification:

What mental image do you get from listening to this song?

Word	Agreement count (N=8)
Fast or chaotic	4
(Battle) field, fighting a war	4
Dark	3
Castle	2

What part of the song influenced your mental picture the most?

Word	Agreement count (N=8)
Background, battle, or sword sounds	6
Fast, chaotic, or busy	5

Would you describe the song you heard as pleasant?

Word	Agreement count (N=8)
Yes	6
No	1
Neutral	1

While looking at a painting:

What sound/song do you imagine when looking at the painting?

Word	Agreement count (N=8)
Chaos or random	5
Battle, fighting, or war sounds	5
Rhythm	2
Marching band, battle drums, or drums	2
Loud, a lot of sounds	2
Trumpets	2
Minor	2

While looking at the painting and hearing the sonification:

Do you think the song is descriptive of the painting?

Word	Agreement count (N=8)
Yes	7
No	0
Somewhat or parts	1

Could anything be added or changed?

Word	Agreement count (N=8)
Could be darker	5
More epic, bombastic	2

Five experts mention that their mental image contains something related to the concepts of “Fast and chaotic”. This is something the painting reflects. Five experts also mention that their mental image contains something related to the concept of a battlefield. The concept of the battlefield describes the painting pretty well. This concept is likely put into the participant’s mental image through the background scene sound as most participants noted that this is what influences their mental picture the most.

Three experts also mention something related to the concept of “Dark”. Dark fits the mood of the painting well but describes some of the colors, like the sky, a bit less. However, the mood of a painting is hard to define objectively as people can have vastly different opinions about the mood of a painting. Therefore someone might argue that the mood of the painting is happy because the person in the middle of the painting is winning the battle, as one expert mentioned.

Because of the instruments used and the timbre of the sound at least two experts mentioned that they got a mental picture of a castle. Although one might connect the battle and horses to the concept of a castle, there is nothing directly related to a castle visible in the painting. Furthermore, the framework does not relate high-level features like objects to the timbre of the sound. Therefore, if the painting contained a castle this connection would be purely accidental.

The majority of the experts described the sonification they listened to as pleasant.

When the experts were looking at the painting they noted that they would add chaos or randomness to the sonification. The sound of a Battlefield is also something the experts would add to their sonification. It is important to note that this question type always came after listening to the sonification made by the framework. Therefore participants could be influenced when coming up with ideas on how to sonify a painting. Something experts mention is the addition of a specific rhythm or marching battle or drums. This idea was not present in the current sonification.

When the painting and the sonification made by the framework were shown at the same time 7 out of 8 participants noted that they found the sonification fitting to the painting. When asked

if they would change anything experts noted mostly that the sound could be darker and also that the sound could be grander.

9.3.2 Alfred Sisley - Snow at Louveciennes



Figure 59: Alfred Sisley - Snow at Louveciennes

https://drive.google.com/file/d/1qWBtLGzfeAmC17L8yIaxwuD88I2A_Ngf/view?usp=sharing

While listening to a sonification:

What mental image do you get from listening to this song?

Word	Agreement count (N=8)
Dark, ominous, evil, or night	5
Castle	2

What part of the song influenced your mental picture the most?

Word	Agreement count (N=8)
Dark or low	3
Not happy	2

Would you describe the song you heard as pleasant?

Word	Agreement count (N=8)
Yes	5
No	3
Neutral	0

While looking at a painting:

What sound/song do you imagine when looking at the painting?

Word	Agreement count (N=8)
Calm, slow, or peaceful	5
Sound of wind	4
Silent, muted sounds or quite	3
Footsteps in snow or movement	3
Cold sounds	2
Minor	2
Sad	2

While looking at the painting and hearing the sonification:

Do you think the song is descriptive of the painting?

Word	Agreement count (N=8)
Yes	5
No	0
Somewhat or parts	3

Could anything be added or changed?

Word	Agreement count (N=8)
Nothing	2

Five experts state that their mental image from this sonification is related to something dark, ominous, evil, or night. Experts stated this was mostly influenced by the dark or low sound and the not happy feeling of the music piece. The person in the center of the painting inhibits dark colors and the white colors of the snow are leaning to gray. Therefore one may interpret the painting as dark. When the person is walking alone through the snow while it's dark the mood of the painting can be interpreted as ominous. This, however, depends on how one interprets the painting. A person going for a happy and calm midnight walk through the snow can also be an interpretation of the painting. The latter has less connection to the mental picture of most experts. Here you can see that creating a general interpretation of a painting in music is difficult because people can have vastly different interpretations of the same painting.

Five of the participants noted that the musical piece was pleasant while three say it's not pleasant. Here some of the participants noted that the unpleasantness came from the dark mood the musical piece conveyed.

When the experts looked at the painting they noted they imagined a sound that was calm, slow, or peaceful. The sounds would be silent, muted, or quiet. As a background sound, they would add the sound of the wind and the footsteps of the person walking in the snow. Because of the snow setting experts also noted they would add sounds that are related to the coldness that the painting portrays. At least two experts noted that they would make the musical piece minor or let it have a sad mood. These last two concepts are somewhat related to what experts experienced while creating their mental picture of the sonification.

Compared to the previous painting there seems to be less relation between the mental picture created by the experts while listening to the sonification and the musical piece imaged while looking at the painting. This notion is also reflected by the expert's opinion of the descriptiveness of the sonification. Five experts say the sonification is descriptive of the painting while three say it's only somewhat descriptive of the painting. Despite three experts finding the sonification somewhat fitting, there is no agreement between all experts on something that could be changed to make the sonification more descriptive.

9.3.3 Enrique Simonet El - barbero del zoco



Figure 60: Enrique Simonet El - barbero del zoco

<https://drive.google.com/file/d/1wxwIPIF-GeBxCanzmZvV1VTs7DBUQqyZ/view?usp=sharing>

While listening to a sonification:

What mental image do you get from listening to this song?

Word	Agreement count (N=8)
Bright, happy, high, excited, playful, summer, sunny, or day	6
Something with people, city, park or eating	5

What part of the song influenced your mental picture the most?

Word	Agreement count (N=8)
People	3
Low and high	2
Melody	2
Energetic	2

Would you describe the song you heard as pleasant?

Word	Agreement count (N=8)
Yes	6
No	0
Neutral	2

While looking at a painting:

What sound/song do you imagine when looking at the painting?

Word	Agreement count (N=8)
Oriental, Arabic, or middle eastern instruments/sounds	6
Flute	3
People speaking	3
Lifelike, daily life	2
Percussion because pots in the background, instruments of cooking gear	2
Regional tuning	2

While looking at the painting and hearing the sonification:

Do you think the song is descriptive of the painting?

Word	Agreement count (N=7)
Yes	3
No	2
Somewhat or parts	2

Could anything be added or changed?

Word	Agreement count (N=7)
More middle eastern sounds/instruments, other instruments	3

While listening to the sonification six experts noted their mental image is related to the concepts of “bright, happy, high, excited, playful, summer, sunny, or day”. Five experts agree on the concepts of “something with people, city, park or eating”. The mental pictures are for most experts influenced by the sound of people in the background. For at least two experts it is mostly influenced by the low and high tones, the melody, or the energy within the musical piece.

Six thought the musical piece was pleasant were two found it neutral.

While looking at the painting most experts would create a musical piece containing Oriental, Arabic, or middle eastern instruments/sounds. Three experts named a more specific instrument to add, namely the flute, and two even talked about using the regional tuning within the musical piece. Besides the instrumentation and tuning, experts would also add a background sound of people speaking, to fit the scene. The sound according to two experts needed to be life-like or represent daily life. The pots in the background of the painting made two experts think of adding percussion sounds to represent the pots.

Although “bright, happy, high, excited, playful, summer, sunny, or day” and “something with people, city, park or eating” seem close to the content of the painting, experts are more focused on the location of the painting. They would create a musical piece containing regional correct instruments and even incorporate the local tuning. This result also shows in the question about the descriptiveness of the painting. Only three experts out of the seven who answered the question find the sonification descriptive of the painting, two find it not descriptive, were two others find it somewhat descriptive. Three experts agree that the current sonification of the painting could use more middle eastern sounds or instruments or other instruments in general.

9.3.4 John Lavery - The Fairy Fountain



Figure 61: John Lavery - The Fairy Fountain

https://drive.google.com/file/d/1M7rk2a1mg-p48iL2IS8PILZgf_FjAw6-/view?usp=sharing

While listening to a sonification:

What mental image do you get from listening to this song?

Word	Agreement count (N=8)
Hard to imagine, weird or difficult	4
Water, waterfall, river, or boat	4
Dark or sad	3
People, talking, voices or speech	3
Feeling of development	2

What part of the song influenced your mental picture the most?

Word	Agreement count (N=8)
Water or rainy sound	3
(Random) melody	2
Dark or low sounds	2
Shifting or jumping sound	2

Would you describe the song you heard as pleasant?

Word	Agreement count (N=8)
Yes	5
No	2
Neutral	1

While looking at a painting:

What sound/song do you imagine when looking at the painting?

Word	Agreement count (N=8)
Water, fountain	6
Darkness, dark timbre, or sad chords	4
Calm, slow	3
Low	2
Mood depending on interpretation	2
Could be happy or sad	2
Talking	2

While looking at the painting and hearing the sonification:

Do you think the song is descriptive of the painting?

Word	Agreement count (N=7)
Yes	5
No	2
Somewhat or parts	0

Could anything be added or changed?

Word	Agreement count (N=7)
Calmer (melody)	3
Background sound softer	2

For this sonification four participants noted that they had difficulty with creating a mental image. Four experts said their mental image contained something related to water, waterfall, river, or boat. This was likely influenced by the water or rainy sound in the background as three experts mention this influenced their mental picture the most. Three experts mentioned that their mental image was dark or sad. This is likely because the dark or low sound two experts said influenced their mental picture the most. Also, three experts mentioned people talking in their mental image and two experts noted a feeling of development. The (random) melody influenced the mental picture of the two experts the most. Two experts also mentioned the shifting or jumping of the sound influenced their mental picture the most. This instability could be an explanation for the difficulty experts had imaging a painting while hearing the sonification.

Five out of eight participants found the musical piece pleasant. Two noted they found it not pleasant and one said it sounded neutral.

Six and thereby most experts would imagine for this painting is the sound of water or a fountain. Four experts would use a dark timbre or sad chords. At least two experts imaged both of these concepts while listening to the sonification. Something the expert did not note in their mental picture is calmness, which is something they would add to their representation of the painting in sound. Two experts would also add the sound of talking in the background. The difficulty to extract the mood of the painting was mentioned by two experts as they stated it could either be happy or sad depending on how the people in the painting are feeling.

The concepts of water, dark, and people talking fit the original painting well despite participants having difficulty creating a mental picture. This difficulty is also not reflected in the descriptiveness of the sonification as five out of seven experts say it's descriptive and only two say it's not. Experts say the current sonification could be calmer and the background sound should be softer.

9.3.5 Claude Monet - Water Lilies

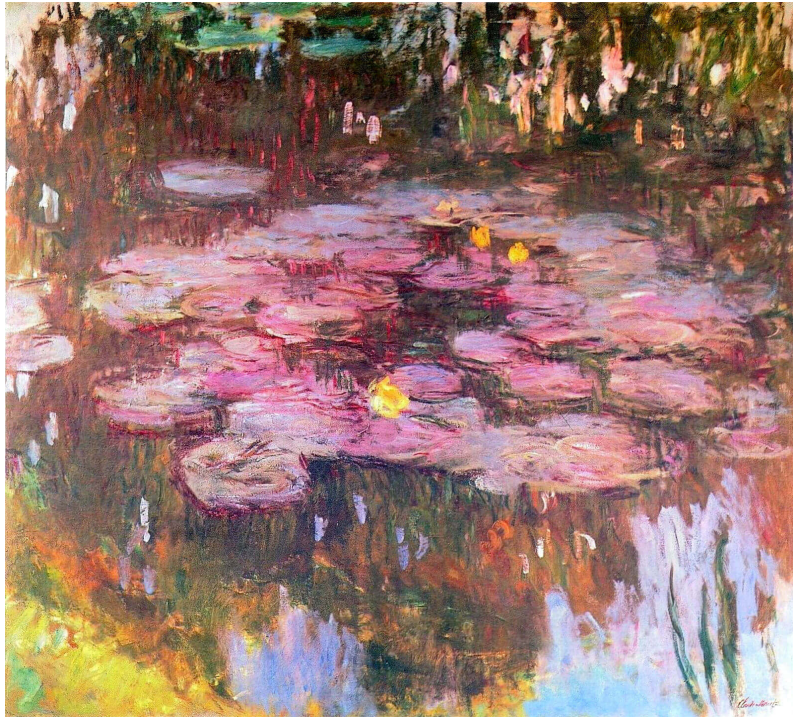


Figure 62: Claude Monet - Water Lilies

<https://drive.google.com/file/d/1sAeT5y7cUBwDfjkw8EP0xYHG1vEQfKZQ/view?usp=sharing>

While listening to a sonification:

What mental image do you get from listening to this song?

Word	Agreement count (N=8)
Water, sea, nature, or boat	8
Not too bright or dark/lower tones	3
Calm	2
Mysterious or surprising	2

What part of the song influenced your mental picture the most?

Word	Agreement count (N=8)
Water sound	7
Horn, trombone, or brass	5
Low tones or bass	4

Would you describe the song you heard as pleasant?

Word	Agreement count (N=8)
Yes	8
No	0
Neutral	0

While looking at a painting:

What sound/song do you imagine when looking at the painting?

Word	Agreement count (N=8)
Water	4
Calm, peaceful, smooth, or slow	4
Happy, cheerful, or warm	3
Frog or water animals	2
Fast or quick	2
Soft	2
Major	2
Bring the reflection into sound	2
Focus on detail not whole	2

While looking at the painting and hearing the sonification:

Do you think the song is descriptive of the painting?

Word	Agreement count (N=7)
Yes	7
No	0
Somewhat or parts	0

Could anything be added or changed?

Word	Agreement count (N=7)
Water sound too moving	4
Remove low tones, remove dark chords	2
Somewhat or parts	0

All experts mention something related to the concepts of “water, sea, nature, or boat”. This part of the mental picture is likely conveyed by the scene sound of water as seven experts say their mental picture was influenced the most by the sound of water. Three participants mentioned that their mental picture was not too bright or they noticed the darker lower tones. Two experts found their mental picture to be calm. Two experts also found their mental picture mysterious or surprising.

Five experts noted that their mental image was influenced the most by the instrument selection of “Horn, trombone, or brass”. Four mentioned that their mental picture was influenced the most by the low tones or bass, which likely influences the concepts of “not too bright or dark/lower tones”

All participants perceived this sonification as pleasant.

Four experts imagine the sound of water while looking at the painting. Something related to the concepts of calm, peaceful, smooth, or slow was also mentioned by four experts. The previously mentioned concepts of water and calmness were both noted by the participant while they created a mental picture of the current sonification. Three experts would create a musical piece that conveys something happy, cheerful, or warm. Two experts would add some sound of a water animal or specifically a frog. One thing that stands out for this painting is that four participants mention that they would create a musical piece that would be calm or slow, two participants also mention they would create something fast or quick. These two concepts could be considered contrasting. This also shows that creating a general sonification can be hard as the interpretation of a painting and the relation to sound is quite subjective. Another thing that stands out is that two experts mentioned that they would add the reflection present in the painting to the sonification. E.g. by inverting the melody at some point in the sonification. Two experts also mentioned that this painting made them more focused on the details within the painting than on the whole.

For this painting, all seven participants said that the sonification was descriptive of the painting. Although the sonification was fitting four experts mentioned that the water sound was too moving for still water and two experts noted that some of the dark sounds could be removed.

9.3.6 William Merritt Chase - The Olive Grove



Figure 63: William Merritt Chase - The Olive Grove

https://drive.google.com/file/d/1qe0NNxdbiVH42-1MkooGzKVGxR7_r21a/view?usp=sharing

While listening to a sonification:

What mental image do you get from listening to this song?

Word	Agreement count (N=8)
Night or dark	5
Forest, jungle, or nature	4
Calm or slow	3
Castle or medieval	2

What part of the song influenced your mental picture the most?

Word	Agreement count (N=8)
Night or dark	5
Forest, jungle, or nature	4
Calm or slow	3
Castle or medieval	2

Would you describe the song you heard as pleasant?

Word	Agreement count (N=8)
Yes	5
No	3
Neutral	0

While looking at a painting:

What sound/song do you imagine when looking at the painting?

Word	Agreement count (N=8)
Slow, calm, relaxing	6
Happy or joyful	5
Nature sounds, spring soundscape	4
Flute	2
High	2
Birds	2
Static scene	2
Leafs in the wind	2

While looking at the painting and hearing the sonification:

Do you think the song is descriptive of the painting?

Word	Agreement count (N=7)
Yes	2
No	4
Somewhat or parts	1

Could anything be added or changed?

Word	Agreement count (N=7)
Too dark	4
Night because of crickets	3
More high tones	2
Less crickets	2
Add birds	2

Five participants noted that the mental image they get from listening to the sonification is related to night or dark. This is a stark contrast to what the painting actually contains. This is likely due to the low and dark tones present in the sonification as three experts mention that low or dark notes mostly influenced their mental image. Experts also mentioned that the cricket sound present in the background made them think of a night scene. The concepts of forest, jungle, or nature fit the painting better and are also likely influenced by the cricket sound in the background. Three experts mentioned that their mental image was calm or slow. These concepts also fit the painting better. Two experts noted that their mental picture contained something related to a castle or medieval times. This is caused by the timbre and instruments present in the sonification as two experts mentioned their mental image is influenced the most by a horn sound. Something interesting to note is that these concepts are brought up more often by experts while none of the paintings in the evaluation dataset contain something related to these concepts.

Of the eight participants, five said the musical piece was pleasant while three participants experienced it as unpleasant.

When the participants were looking at the painting, six noted that they imagined a slow, calm, relaxing sound and would add a nature sound to the background. This corresponds with the mental image some experts had while hearing the current sonification of this painting. Five experts would create a "Happy or joyful" sound to go with the painting. This is in contrast to what some participants imagined while listening to the current sonification. Two experts noted they specifically would add a flute as an instrument and birds or leaves in the wind as background. Two participants also noted that the musical piece should be played in a high register.

Only two experts found this sonification fitting to the painting while four do not. One participant found it only somewhat fitting. This is likely due to the dark tone the sonification presents and the crickets in the background that convey a night scene. Experts noted that the sonification was too dark to be fitting to the painting and the crickets made it too much like a night scene. This can also be seen in the creation of the mental image by experts as they tended to imagine a dark night scene.

9.3.7 Valentin Serov - Iphigenia in Tauris



Figure 64: Valentin Serov - Iphigenia in Tauris

<https://drive.google.com/file/d/1t20LsuMxL4bSb0X0efP9qiDI6E7pWtc/view?usp=sharing>

While listening to a sonification:

What mental image do you get from listening to this song?

Word	Agreement count (N=8)
Waves, beach, forest, or nature	4
Medieval	3
Fantasy	2

What part of the song influenced your mental picture the most?

Word	Agreement count (N=8)
Melody or note pattern	2
White noise	2

Would you describe the song you heard as pleasant?

Word	Agreement count (N=8)
Yes	7
No	1
Neutral	0

While looking at a painting:

What sound/song do you imagine when looking at the painting?

Word	Agreement count (N=8)
Wave, sea sound	7
Calm, low tempo	3
Sad	2
Grand, big feeling	2
High	2

While looking at the painting and hearing the sonification:

Do you think the song is descriptive of the painting?

Word	Agreement count (N=8)
Yes	8
No	0
Somewhat or parts	0

Could anything be added or changed?

Word	Agreement count (N=8)
Melody is all over the place	2
Wave sound can be more wavy	2

Four experts mentioned that the mental image they got from listening to the sonification was related to the concepts of waves, beach, forest, or nature. While forest is also a nature-related concept, this is not what the painting contains. The "Waves and beach" describe the painting pretty well. Three experts said their mental image was related to medieval times and two noted that it was related to fantasy. This is likely caused by the timbre of the sound as mentioned previously. Two experts mentioned that the melody or note pattern influenced their mental image the most. Also, two experts mentioned that the white noise in the background influenced it the most.

Seven experts found the sonification pleasant while one found it unpleasant.

While looking at the painting seven experts noted that they imagine a wave or sea sound. This fits with the wave sound present in the current sonification. Three experts would create a musical piece that is calm or low tempo and two experts would let it represent a sad emotion. It is good to note that not all experts considered the mood of the painting as sad, as one expert explicitly mentioned it to be a happy context. Two experts would create a grand musical piece. Also, two experts would include mostly high notes or sounds.

All the experts found this sonification fitting to the painting. Although one expert said that all sonifications are quite ambient and ambient fits this painting well. Therefore, it is good to note that this result can only say that the 4Th model works well for this particular painting and does not tell anything about the performance of different paintings. Besides all the experts finding the sonification descriptive of the painting two experts noted that the melody was too random. Two experts also noted that the wave sound in the background could sound more like actual waves.

9.3.8 Paul Delvaux - The Viaducto



Figure 65: Paul Delvaux - The Viaducto

https://drive.google.com/file/d/10jxghKE2Gn3D_5MCwP8SNFAqN70tgFC2/view?usp=sharing

While listening to a sonification:

What mental image do you get from listening to this song?

Word	Agreement count (N=8)
Waves, beach, forest, or nature	4
Medieval	3
Fantasy	2

What part of the song influenced your mental picture the most?

Word	Agreement count (N=8)
Melody or note pattern	2
White noise	2

Would you describe the song you heard as pleasant?

Word	Agreement count (N=8)
Yes	7
No	1
Neutral	0

While looking at a painting:

What sound/song do you imagine when looking at the painting?

Word	Agreement count (N=8)
Wave, sea sound	7
Calm, low tempo	3
Sad	2
Grand, big feeling	2
High	2

While looking at the painting and hearing the sonification:

Do you think the song is descriptive of the painting?

Word	Agreement count (N=8)
Yes	8
No	0
Somewhat or parts	0

Could anything be added or changed?

Word	Agreement count (N=8)
Melody is all over the place	2
Wave sound can be more wavy	2

When describing their mental image two experts mentioned the concepts of the night sky or blue sky which describes the painting. Two experts mentioned storm or wind to be present in their mental image which does not describe the painting well. Three participants said the biggest influence on their mental image is the chords of the sonification.

Two experts described the painting as pleasant, two as unpleasant, and four as neutral.

When experts described the sound they saw fitting for the painting, six experts mentioned the concepts of calm or slow. Four experts would incorporate the rhythm of a train within the sonification. Also, four participants would add the sound of a night scene. Three experts would add the sound of a train or a railroad as a background sound. Two experts mentioned the explicit use of a piano.

Five experts found the sonification descriptive of the painting, one found it not descriptive and two experts found it somewhat descriptive. Three experts mentioned that the sonification was all over the place. This might also explain the low agreement count in when imagining a mental picture for the sonification.

9.4 Two paintings, one sonification

9.4.1 Charge of the scots greys at waterloo & William Merritt Chase - The Olive Grove



Figure 66: Charge of the scots greys at waterloo



Figure 67: William Merritt Chase - The Olive Grove

Sonification experts heard (right):

<https://drive.google.com/file/d/1fmmPPaHqPuAHjjC8PKE7npEFWhUx8lxF/view?usp=sharing>

Could you choose one of the two paintings you find best fitting to the musical piece?

Word	Agreement count (N=4)
Correct	1
Incorrect	0
Unsure	3

Could you explain your choice?

Word	Agreement count (N=4)
Background or nature sound	2
Mood more left background more right	2

What aspects of the song or the painting stood out or influenced your choice the most?

Word	Agreement count (N=4)
Background, nature, or insect sounds	3

When confronted with the two paintings above and the sonification of the painting on the right, most experts were unsure to which painting the sonification they heard accurately belonged to. The experts noted that the mood of the sonification was better fitting to the left painting than to the right, but were inclined to choose the right painting because of the background noise. The background noise consisted of wind and crickets.

9.4.2 Valentin Serov - Iphigenia in Tauris & Claude Monet - Water Lilies



Figure 68: Valentin Serov - Iphigenia in Tauris

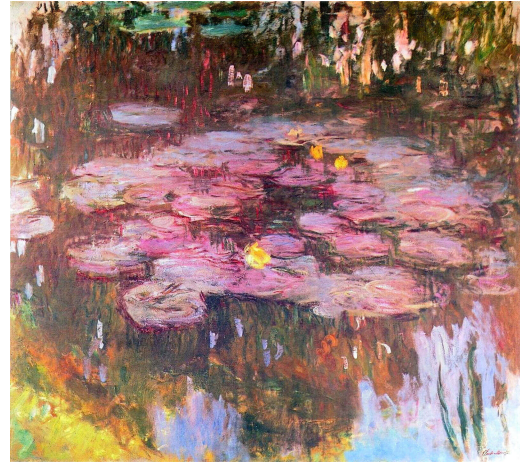


Figure 69: Claude Monet - Water Lilies

Sonification experts heard (left):

<https://drive.google.com/file/d/1pgTgcapfTGG8R40hwUj7PT4DEcLyEG5j/view?usp=sharing>

Could you choose one of the two paintings you find best fitting to the musical piece?

Word	Agreement count (N=4)
Correct	4
Incorrect	0
Unsure	0

Could you explain your choice?

Word	Agreement count (N=4)
Background sound, water, or the sound of the sea	4

What aspects of the song or the painting stood out or influenced your choice the most?

Word	Agreement count (N=4)
Water or wave sounds	3

For this pair of paintings, every expert chose the left painting, which corresponded to the sonification they heard. All experts noted that the background sound was part of the reason they chose the left painting and three say it even influenced their choice the most.

9.4.3 Alfred Sisley - Snow at Louveciennes & Enrique Simonet - El barbero del zoco



Figure 70: Alfred Sisley - Snow at Louveciennes



Figure 71: Enrique Simonet - El barbero del zoco

Sonification experts heard (right):

<https://drive.google.com/file/d/1vga5NMscAQzixitT0zwcw03xFuq06T-Qd/view?usp=sharing>

Could you choose one of the two paintings you find best fitting to the musical piece?

Word	Agreement count (N=4)
Correct	3
Incorrect	1
Unsure	0

Could you explain your choice?

Word	Agreement count (N=4)
People talking, movement	3

What aspects of the song or the painting stood out or influenced your choice the most?

Word	Agreement count (N=4)
Background sound	2

For this pair of paintings, three out of four experts chose the correct painting choice while hearing the sonification of the painting on the right. Most experts noted their decision was based on the movement of people talking. This is most likely caused by the background sound as it contained the sound of people talking. Two experts also noted that the background sound influenced their choice the most.

9.4.4 Paul Delvaux - the viaducto & John Lavery - The Fairy Fountain



Figure 72: Paul delvaux - the viaducto

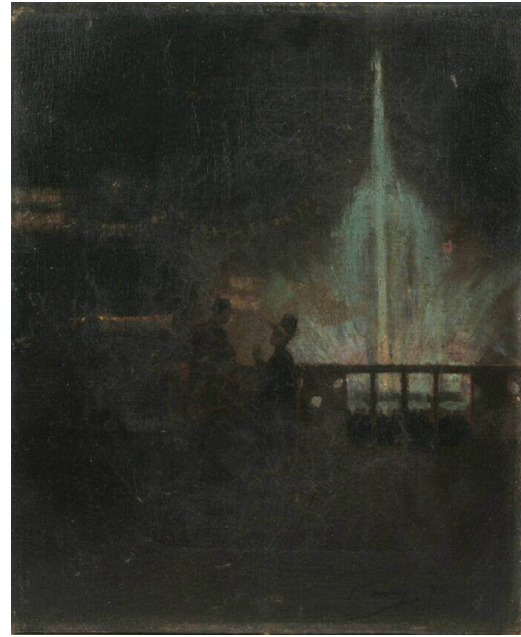


Figure 73: John Lavery - The Fairy Fountain

Sonification experts heard (left):

https://drive.google.com/file/d/12FeeKrAUK_OxxwTkrs_AlfD3phvFNC-M/view?usp=sharing

Could you choose one of the two paintings you find best fitting to the musical piece?

Word	Agreement count (N=4)
Correct	4
Incorrect	0
Unsure	0

Could you explain your choice?

Word	Agreement count (N=4)
Train, mechanical, background sound	3
Difficult, doubt	2

What aspects of the song or the painting stood out or influenced your choice the most?

Word	Agreement count (N=4)
Sound of the train	2
Difficult, doubt	2

For this pair of paintings, all experts chose the correct painting while listening to the sonification of the left painting. One interesting thing to note is that while all experts made the correct choice, two experts noted that they found it difficult to choose or were doubtful about their choice. In this sonification, the most significant influence was the background sound as three experts named something related to the train, mechanical, or background sound while explaining their choice and two experts told that the sound of the train influenced their decision the most.

9.5 General feedback

When looking at the results people tend to extract the setting of the painting mostly via the added background sound of the sonification. This is logical and expected as this is a direct link to the setting of the painting. People tend to extract the mood of the painting more through the chord progression and the melody. The emotions extracted were mostly limited to calm or chaotic or happy or sad. This is also as expected as only the darkness, lightness, colors, and object amount are portrayed by the chord and melody progression. One element of the sonification that seldom appeared in the explanation of the mental picture of experts is color. This is interesting because, when experts were asked about how they would implement a sonification method, multiple experts talked about extracting color to create a sound, however, when faced with a sonification their explanation of their mental image rarely contained specific colors. Therefore it is likely that people think turning color into a sound is an intuitive way of encoding visual information into sound, but when extracting information from sound, this method looks less intuitive. Experts tended to talk more about the setting or objects within their mental picture, or the emotion the sonification brought them. Another explanation for this could be that the 4Th model does not put enough emphasis on the color present in the painting, therefore the focus tends to be on other parts of the painting. However, when experts were asked what sound they would imagine for a painting they hardly mentioned colors to be linked to specific sounds or instruments. Rather they link specific objects or the setting of the painting to a timbre. This shows the most in Figure 60 where experts want to use regional instrumentation or even tuning to represent the painting. If experts commented on the color, the reaction was mostly about the lightness of the color, i.e. light or dark. When a painting was darker experts tended to note that they would make a progression in the minor scale, although this can also be the case for light paintings with a possible sad interpretation Figure 64. Experts also tended to add explicit sounds to describe the setting, such as sounds of objects present in the scene or sounds you could hear in the context of the painting.

The experts were asked if they found the sonification for a painting descriptive of the painting itself and if there was something they would change to make it more descriptive. One feedback point that was given for multiple paintings was that the sound was too dark for the content of the painting, meaning there were low notes played while the painting only contained bright colors. Another point of feedback is about the randomness of the melody. Some experts noted that the melody had no musical structure and thereby gave the feeling that the melody was random. Other points of feedback are related to bringing more of the setting or high-level features to the sonification. Adding more sounds related to the specific objects in the painting, or the timbre and rhythm of the sound more related to the setting, are a couple of feedback points that were mentioned.

One challenge that became clear during the evaluation is the problem of the subjective interpretation of paintings and sonifications. Therefore creating a sonification framework that creates sonifications that are descriptive for every individual seems impossible, but a framework that aims to create descriptive sonifications for the general public seems feasible.

While doing the interviews some things stood out. Experts noted that after hearing the sonifications for a while everything started to sound very similar and got a bit tired of listening. This problem can probably be attributed to a lack of diversity between the sonifications. Related to the former experts found the timbre of the sound related to medieval times while no paintings with this context were present within the interview.

10 Conclusion

With this research we try to answer the question of “How can high-level visual features present in paintings be incorporated in an automated and pleasant painting sonification method.”. To make the question more comprehensible the question has been divided into three sub-questions.

- How can existing sonification methods contribute to the automation of painting sonification?
- How can a sonification pipeline be created to incorporate high-level features extracted from paintings?
- How will the overall quality and the value of the addition of high-level features to the sonification be validated?

To answer the first sub-question literature research has been conducted to see what the current methods of sonification are and how they are to value in the creation of an automatic framework for painting sonification. Existing work used mainly low-level features present in the visual space of paintings, such as color and edge. The work of Rector et al. [32] introduced the use of high-level features in painting sonification. With these concepts in mind, the research set out to answer the second sub-question. To answer this question a framework has been created to test the feasibility of a framework that can automatically sonify painting using low and high-level features. For the sonification of low-level features, this research takes inspiration from Cavaco et al.[4] by linking HSV values to sound properties but also takes inspiration from Polo et al.[29] of a noncontinuous linking of color to piano notes to create a more harmonious sound. Furthermore, a similar idea as used by Kabisch et al.[19] was implemented for edge information. The edge, if present in a segment, is used to change the timbre of the sound with the process of waveshaping. To convey the high-level feature of the scene present in a painting, scene extraction is used to add the sound of a scene to the sonification. Object segmentation is used to create the structure of the sonification based on the segments present in a painting.

With the previous design ideas in mind, four models have been created. The first model extracted the dominant color from a painting divided in square segments. The HSV of the dominant color and edge information were used to create a chord, melody, and timbre for a specific segment. To be able to influence the timbre based on edge information a wavetable synthesis method was implemented. The wavetable synth existed of three waveforms, one sine wave, one waveform created from the histogram of the painting, and one saw wave. These waveforms were chosen to convey the roughness of a segment and to create a unique timbre for each painting. To create a sense of space the location of a segment was used to create the panning of a segment within the sonification. The more a segment was located on the left or right of the painting the more the segment would sound on the left or right. To navigate between different segments saliency was used to create an ordering where the most salient segment was sonified first and the least salient segment was sonified last. However, the results of the first model encountered some problems. The segmentation of the paintings into squares felt unnatural and the diversity between different sonifications was small.

Therefore, a second model was created to counteract these problems. The segmentation was now based on objects instead of squares to create a more natural segmentation. This new segmentation also affected the creation of panning and the melody. In an attempt to create a more diverse sonification the second model implemented a FM synthesis method. However, the second model did not solve the diversity problem.

Therefore a third model was created, which instead of using synthesis, was producing sounds of existing instruments. Each hue was assigned its own unique instrument in this model. While the diversity of the timbre increased, the sound of the sonification became flatter. Also, the chords of the model stayed the same, therefore this model still inhibited the problem of diversity.

The fourth and last model consisted of a combination of the second and third models on the audio generation level. This was done to keep the diversity of the timbre of the third model, but combat the flat sound it created. To create a more diverse composition Inner scaling was created and the number of objects was an influence on the note duration. Inner Scaling made a scaling based on the visual information present within a painting instead of using a static scaling for all paintings. The last model exhibited the most diversity, however, this diversity shows mostly within the sonification itself and not between paintings. Although there are still improvements that can be made with the former models, the research produced a framework that could produce sonifications including low and high-level features automatically.

To see what problems still exist and what improvements can be made the research tries to answer the last question with an evaluation. Therefore, an experiment has been set up where experts were interviewed to find the strengths and weaknesses of the last model and discover future possibilities. Most experts tended to attribute the descriptiveness of the sonification of the painting to the scene sound present. The mood of the painting was mostly conveyed by the chords of the sonification. According to experts, they would use the color of a painting to turn it into sound, but rarely follow up on this idea when they are asked to turn a painting into sound. The absence of mentioning color while listening to the sonifications could be because the use of color is a weakness in the design of the last model or because the use of color is not as intuitive as toughed when turning paintings into sound. If the former is the case the model could improve by putting more emphasis on color, however if the later is the case then putting more emphasis on the color would not give an increase performance. Furthermore, while looking at paintings, most suggestions by experts on how to turn a painting into sound were based on including some sort of high-level feature in the link of the visual space to the auditory space.

One over arching problem of the framework was the lack of diversity between sonifications, this meant that the framework lost its descriptiveness over time and became harder to listen to because of listening fatigue. Another problem with the concept of painting sonifications is that creating a sonification that is descriptive for every individual seems impossible, but creating descriptive sonifications for the general public seems possible.

11 Discussion and future work

One of the problems this framework encounters is the diversity between sonifications. Because of this, it is hard to differentiate between paintings by only listening to their sonifications. It also has the negative effect of getting easily tired of listening to the sonifications. There are a couple of ideas to address this issue. More diversity could be created by the addition of more scales. The current solution only uses the Major and Minor scales for light and dark paintings respectively. A finer link between the brightness and scale could be used when more scales are available. Another approach in line with expert ideas could be to link more high-level sound properties and content to high-level features present in the painting. One option could be to link a specific object to the timbre of a sound, e.g. a clear and calm sky is composed of long chords and a soft sound, whereas a gun in a battle could be composed of quick and short notes with a harsh sound. Besides the specific objects influencing the timbre and composition, the whole scene could also be used to influence the timbre more specifically. Overall experts were more focused on high-level features present in a painting, rather than thinking about the colors and how they would sound. One good example is Figure 60 where experts go as far as to use regional instrumentation or tuning to represent the painting in the auditory space.

Another problem, which can be seen as a sub-problem of diversity, is the lack of musical cohesion within the sonifications. Currently, a segment that is played does not take into account the previous or next played segment in the sonification, thereby making the musical relationship between the segments up to chance, meaning sometimes one segment can play well with another segment, while sometimes there isn't any relation to be found. This can make the sonifications sound random and

therefore not diverse, making them hard to listen to over a long period of time. A way to solve this issue would be to make sure the chords and melodies of segments take their neighboring segments into account, therefore making the whole sonification a coherent musical piece. Also, the current way of creating the melody of a segment is by taking the color of all segments, thereby the notes do not take neighboring notes into account, therefore creating a seemingly random melody. One way to counteract this randomness would be to create predefined melodies that could convey the emotion of the painting. E.g. when a painting consists of bright colors a predefined happy melody could play for every segment. To make sure the melody does not sound too repetitive a couple of notes could be randomized.

Currently, the framework only works on a selected set of paintings. This is largely due to the fact that it is still difficult to extract high-level features from paintings, such as scene and object segmentation. While the above improvements of the framework suggest more high-level features to be incorporated, no suggestion is made on how to do so. One solution could be to look at more available data about the painting than only its visual aspect, such as text information written about the painting. One could extract the year and location the painting represents and find fitting instrumentation. Or emotion detection on a description of the painting could be used to get a general emotion within the music.

Currently, the research states that the creation of a sonification for the general public seems possible. If a sonification for the general public is possible can not be said with certainty. This is because of a limitation on the type of evaluation used in this research. The evaluation done created a small sample size, therefore nothing can be said over the general public. Therefore a larger evaluation could be set up to create a better sense of the descriptiveness of the sonifications to the general public. To realize this a website could be created. On this website, participants should be able to answer questions regarding the descriptiveness of the sonification based on the painting. To keep the evaluation easy to consume, only one question could be asked at a time. Also, participants should be able to continue for as long as they want, meaning they could answer one question or as many as they wanted to make the evaluation easy to enter. This research tried to answer the question: "How can high-level visual features present in paintings be incorporated in an automated and pleasant painting sonification method." by using a bottom-up research method and creating a framework based upon existing work. However, during evaluation, it became clear that a bottom-up approach did not suit this research as well as first thought. This could be since existing work tried to solve a different problem than this research, by creating sonification methods to convey information to the visually impaired. Because of their aim, automation or pleasantness was not a priority. Also, previous research included a training phase in their evaluation, something that was not present in the evaluation of this research. This can also be included in a future evaluation to see if training improves the descriptiveness of the sonifications.

Future research into this topic would benefit from a top-down approach where the focus is more on how people would turn a painting into a musical piece and what information people tend to extract during this process. This could be done by creating a website where people could describe how they would turn a painting into sound and find common concepts in their descriptions. This information can give insight into whether it is possible to create sonifications of paintings for the general public given existing technology. If the former seems possible, a bottom-up approach can be used to use the gathered information to create a framework more specific to the problem of painting sonification for the extension of the art experience. This framework can be evaluated on a large scale, as described in the former paragraph, to see if the descriptiveness of the sonifications improves and is in line with the information gathered in the first step.

References

- [1] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. “Soundnet: Learning sound representations from unlabeled video”. In: *arXiv preprint arXiv:1610.09001* (2016).
- [2] Andrew J Bremner et al. ““Bouba” and “Kiki” in Namibia? A remote culture make similar shape–sound matches, but different shape–taste matches to Westerners”. In: *Cognition* 126.2 (2013), pp. 165–172.
- [3] Jean-Pierre Briot and Francois Pachet. “Music generation by deep learning-challenges and directions”. In: *arXiv preprint arXiv:1712.04371* (2017).
- [4] Sofia Cavaco et al. “Color sonification for the visually impaired”. In: *Procedia Technology* 9 (2013), pp. 1048–1057.
- [5] Jun Dong Cho et al. “Sound Coding Color to Improve Artwork Appreciation by People with Visual Impairments”. In: *Electronics* 9.11 (2020), p. 1981.
- [6] Kevin Crowston. “Amazon Mechanical Turk: A Research Tool for Organizations and Information Systems Scholars”. In: *Shaping the Future of ICT Research. Methods and Approaches*. Ed. by Anol Bhattacharjee and Brian Fitzgerald. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 210–221. ISBN: 978-3-642-35142-6.
- [7] Jia Deng et al. “Imagenet: A large-scale hierarchical image database”. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.
- [8] Prafulla Dhariwal et al. “Jukebox: A Generative Model for Music”. In: *arXiv preprint arXiv:2005.00341* (2020).
- [9] Jesse Engel et al. “Neural audio synthesis of musical notes with wavenet autoencoders”. In: *International Conference on Machine Learning*. PMLR. 2017, pp. 1068–1077.
- [10] Christopher Fassinige, Claudia Cecconi Marcotti, and Elliot Freeman. “A deafening flash! Visual interference of auditory signal detection”. In: *Consciousness and cognition* 49 (2017), pp. 15–24.
- [11] Eduardo Fonseca et al. “FSD50k: an open dataset of human-labeled sound events”. In: *arXiv preprint arXiv:2010.00475* (2020).
- [12] Steven P Frysinger. “A brief history of auditory data representation to the 1980s”. In: Georgia Institute of Technology. 2005.
- [13] MM El-Gayar, H Soliman, et al. “A comparative study of image low level feature extraction algorithms”. In: *Egyptian Informatics Journal* 14.2 (2013), pp. 175–181.
- [14] Ross Girshick. “Fast r-cnn”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1440–1448.
- [15] Ross Girshick et al. “Rich feature hierarchies for accurate object detection and semantic segmentation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 580–587.
- [16] Junwei Han et al. “Advanced deep-learning techniques for salient and category-specific object detection: a survey”. In: *IEEE Signal Processing Magazine* 35.1 (2018), pp. 84–100.
- [17] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.
- [18] Eric Heep and Ajay Kapur. “Extracting visual information to generate sonic art installation and performance”. In: *Proceedings of the 21st International Symposium on Electronic Art*. Vancouver, Canada. 2015.
- [19] Eric Kabisch, Falko Kuester, and Simon Penny. “Sonic panoramas: experiments with interactive landscape image sonification”. In: *Proceedings of the 2005 international conference on Augmented tele-existence*. 2005, pp. 156–163.

- [20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “Imagenet classification with deep convolutional neural networks”. In: *Advances in neural information processing systems 25* (2012), pp. 1097–1105.
- [21] Teng Li et al. “Contextual bag-of-words for visual categorization”. In: *IEEE Transactions on Circuits and Systems for Video Technology* 21.4 (2010), pp. 381–392.
- [22] Wei Liu et al. “Ssd: Single shot multibox detector”. In: *European conference on computer vision*. Springer. 2016, pp. 21–37.
- [23] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. “TUT database for acoustic scene classification and sound event detection”. In: *2016 24th European Signal Processing Conference (EUSIPCO)*. IEEE. 2016, pp. 1128–1132.
- [24] Annamaria Mesaros et al. *TUT acoustic scenes 2017, development dataset*. 2017.
- [25] Edoardo Michelsoni et al. “INTERACTIVE PAINTING SONIFICATION USING A SENSOR-EQUIPPED RUNWAY”. In: (2017).
- [26] Aaron van den Oord et al. “Wavenet: A generative model for raw audio”. In: *arXiv preprint arXiv:1609.03499* (2016).
- [27] Karol J Piczak. “ESC: Dataset for environmental sound classification”. In: *Proceedings of the 23rd ACM international conference on Multimedia*. 2015, pp. 1015–1018.
- [28] Irwin Pollack and Lawrence Ficks. “Information of elementary multidimensional auditory displays”. In: *The Journal of the Acoustical Society of America* 26.2 (1954), pp. 155–158.
- [29] Antonio Polo and Xavier Sevillano. “Musical Vision: an interactive bio-inspired sonification tool to convert images into music”. In: *Journal on Multimodal User Interfaces* 13.3 (2019), pp. 231–243.
- [30] Ashish Ranjan, Varun Nagesh Jolly Behera, and Motahar Reza. “Using a Bi-directional LSTM Model with Attention Mechanism trained on MIDI Data for Generating Unique Music”. In: *arXiv preprint arXiv:2011.00773* (2020).
- [31] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. “Generating diverse high-fidelity images with vq-vae-2”. In: *Advances in Neural Information Processing Systems*. 2019, pp. 14866–14876.
- [32] Kyle Rector et al. “Eyes-free art: exploring proxemic audio interfaces for blind and low vision art engagement”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1.3 (2017), pp. 1–21.
- [33] Joseph Redmon and Ali Farhadi. “YOLOv3: An Incremental Improvement”. In: *arXiv* (2018).
- [34] Shaoqing Ren et al. “Faster r-cnn: Towards real-time object detection with region proposal networks”. In: *arXiv preprint arXiv:1506.01497* (2015).
- [35] Christopher Short. *The art theory of Wassily Kandinsky, 1909-1928: the quest for synthesis*. Peter Lang, 2010.
- [36] Karen Simonyan and Andrew Zisserman. “Very deep convolutional networks for large-scale image recognition”. In: *arXiv preprint arXiv:1409.1556* (2014).
- [37] Dan Stowell et al. “Detection and classification of acoustic scenes and events”. In: *IEEE Transactions on Multimedia* 17.10 (2015), pp. 1733–1746.
- [38] Christian Szegedy et al. “Going deeper with convolutions”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9.
- [39] Roberto Vaz, Diamantino Freitas, and António Coelho. “Blind and Visually Impaired Visitors’ Experiences in Museums: Increasing Accessibility through Assistive Technologies.” In: *International Journal of the Inclusive Museum* 13.2 (2020).

- [40] Yuxin Wu et al. *Detectron2*. <https://github.com/facebookresearch/detectron2>. 2019.
- [41] Jianxiong Xiao et al. “Sun database: Large-scale scene recognition from abbey to zoo”. In: *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE. 2010, pp. 3485–3492.
- [42] Lisha Xiao, Qin Yan, and Shuyu Deng. “Scene classification with improved AlexNet model”. In: *2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*. IEEE. 2017, pp. 1–6.
- [43] Tsubasa Yoshida et al. “EdgeSonic: image feature sonification for the visually impaired”. In: *Proceedings of the 2nd Augmented Human International Conference*. 2011, pp. 1–4.
- [44] Bolei Zhou et al. “Places: A 10 million image database for scene recognition”. In: *IEEE transactions on pattern analysis and machine intelligence* 40.6 (2017), pp. 1452–1464.

A Appendix: Dataset filter stages graphs

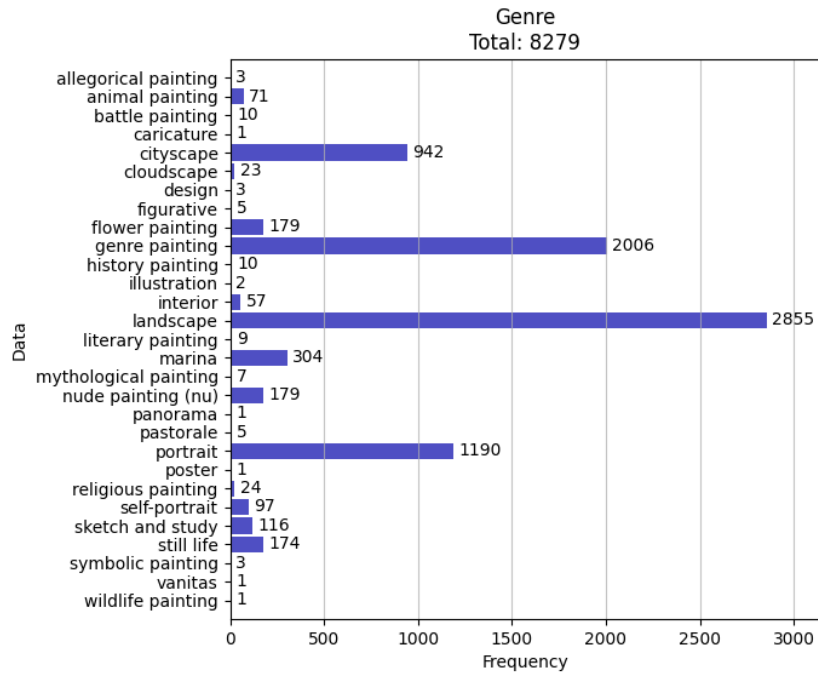


Figure 74: Genres in Painter by Numbers dataset after filtering on Impressionism

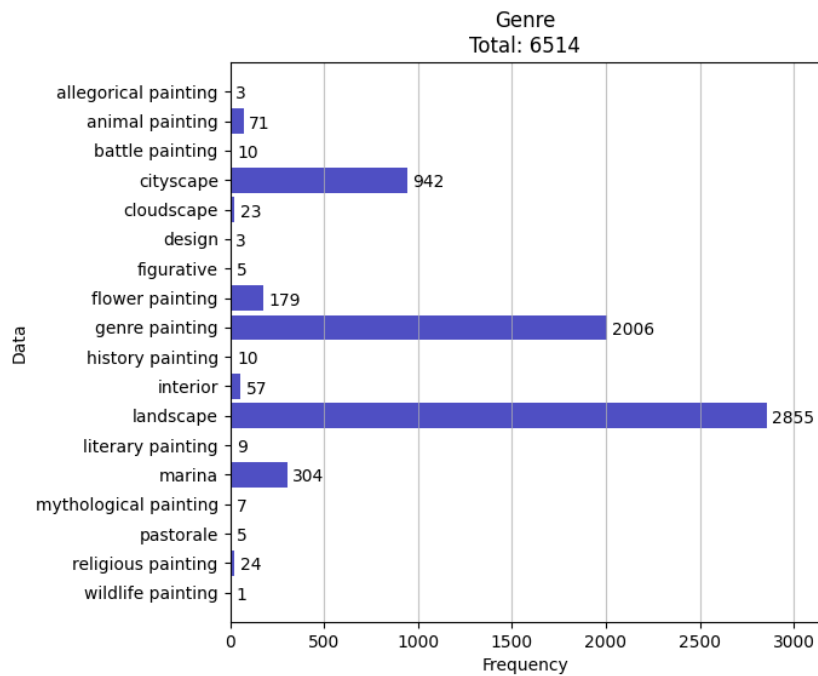


Figure 75: Genres in Painter by Numbers dataset after omitting genres

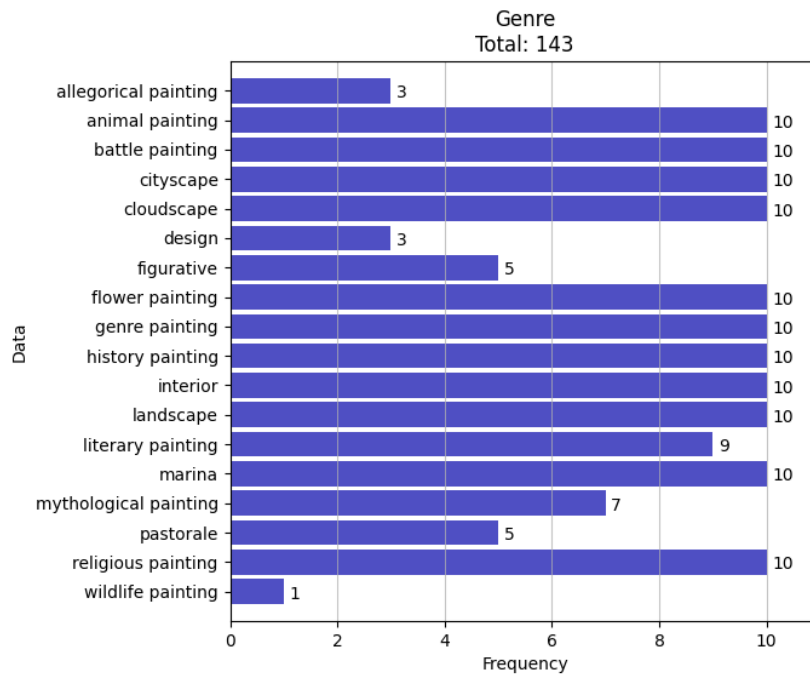


Figure 76: Genres in Painter by Numbers dataset after randomly picking a maximum of 10 paintings per genre

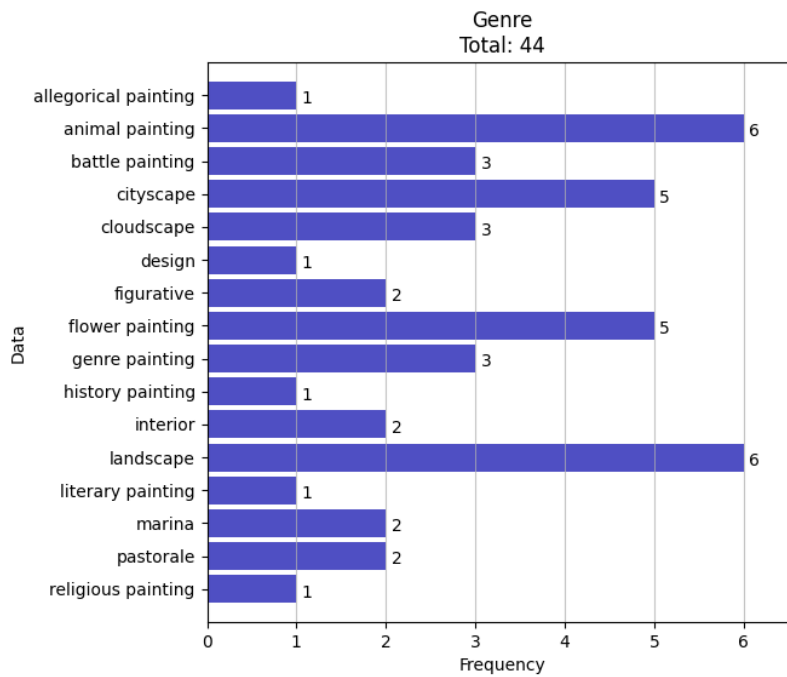


Figure 77: Genres in Painter by Numbers dataset after omitting incorrect scene detection

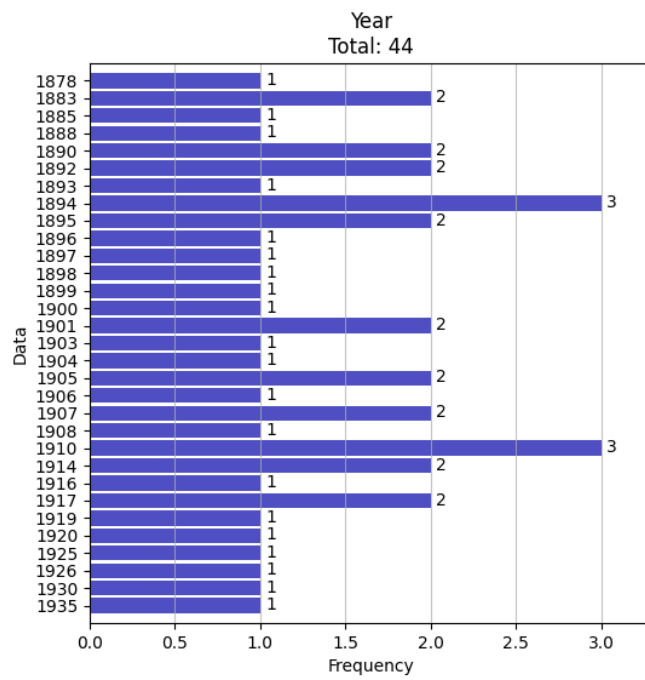


Figure 78: Creation years of paintings in Painter by Numbers dataset after omitting incorrect scene detection

B Appendix: Paintings from the dataset



Figure 79: Charge of the scots greys at waterloo

Sonifications:

Model 1: https://drive.google.com/file/d/1q_tsxcHyyOPGYMH0_G4utFi7qL6jF5pJ/view?usp=sharing

Model 2: <https://drive.google.com/file/d/1V17G8AjTemIxaPMat353u16NwFxp30/view?usp=sharing>

Model 3: <https://drive.google.com/file/d/1oa7RAWUsioBtWzVSLKmp4iBqT-eeWpu8/view?usp=sharing>

Model 4: <https://drive.google.com/file/d/1iMBV5mPiJ17YLgzEbcIvuP0r7d4sAGsL/view?usp=sharing>



Figure 80: Alfred Sisley - Snow at Louveciennes

Sonifications:

Model 1: https://drive.google.com/file/d/1cwf8M0chmGo4jyPX1Hho_po0yzBcNwJ/view?usp=sharing

Model 2: https://drive.google.com/file/d/1uAy_5Q1SYSUtKqfJuwQawuzQVNwn0IWF/view?usp=sharing

Model 3: https://drive.google.com/file/d/1-Tp1GFQGhEKP8zu1AI0b3J_ejQKjuHYy/view?usp=sharing

Model 4: https://drive.google.com/file/d/1qwBtLGzfeAmC17L8yIaxwuD88I2A_Ngf/view?usp=sharing



Figure 81: Enrique Simonet El - barbero del zoco

Sonifications:

Model 1: <https://drive.google.com/file/d/1RqNr7XNs4TxbQh-0tzUpUYN-vSKOBUgZ/view?usp=sharing>

Model 2: <https://drive.google.com/file/d/1bCJpt6KeS8vf4FshIZLGSm4n4eawzws/view?usp=sharing>

Model 3: <https://drive.google.com/file/d/1PXG31zP80ZBtM3Z4S0dbyIBJ3DCGVFEX/view?usp=sharing>

Model 4: <https://drive.google.com/file/d/1wxwIPIF-GeBxCanzmZv1VTs7DBUQqyZ/view?usp=sharing>



Figure 82: John Lavery - The Fairy Fountain

Sonifications:

Model 1: <https://drive.google.com/file/d/1EDGEn9W8vtf6FYs661Zo9ShAwjMZvCIX/view?usp=sharing>

Model 2: <https://drive.google.com/file/d/1tHwkuvEpxYrgvubufoiPTkU0-ovBbJD0/view?usp=sharing>

Model 3: <https://drive.google.com/file/d/1ZAE5Y2DniwAIwIPFETzkcE6q2qyxZM4s/view?usp=sharing>

Model 4: https://drive.google.com/file/d/1M7rk2a1mg-p48iL2IS8PILZgf_FjAw6-/view?usp=sharing

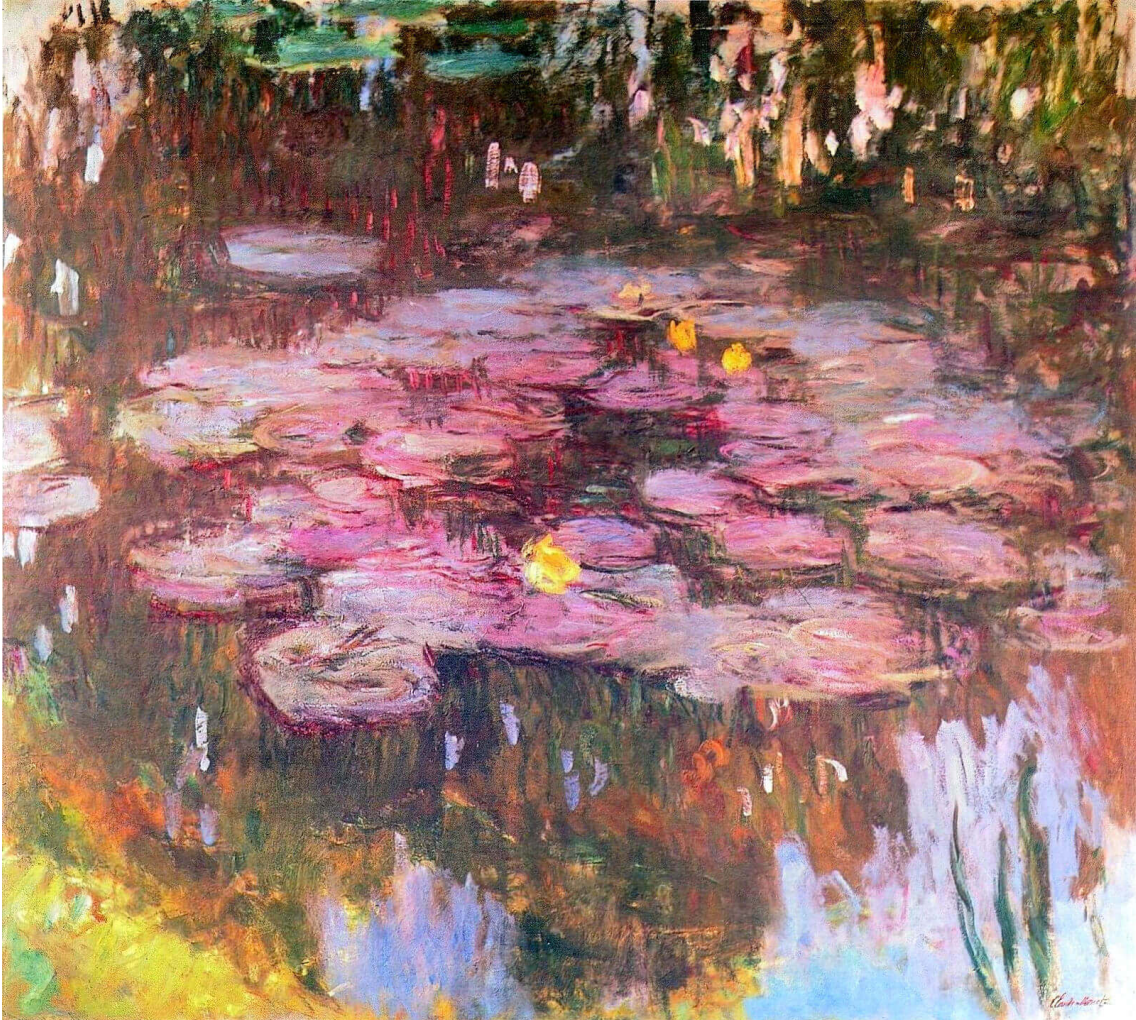


Figure 83: Claude Monet - Water Lilies

Sonifications:

Model 1: https://drive.google.com/file/d/1_nNpeTcIWsR6xAfwkY6c-KjIv6iV0r8/view?usp=sharing

Model 2: https://drive.google.com/file/d/14k3aiRL30RB2YjcxwNM3YJxrSJ668Y_/view?usp=sharing

Model 3: https://drive.google.com/file/d/1olhaA9ne-r89MkGXuQ_XITGyyLpSH4pA/view?usp=sharing

Model 4: <https://drive.google.com/file/d/1sAeT5y7cUBwDfjkw8EP0xYHG1vEQfKZQ/view?usp=sharing>



Figure 84: William Merritt Chase - The Olive Grove

Sonifications:

Model 1: <https://drive.google.com/file/d/1oYpudnBmpvP9nIBn3fkygIEBc5JSSm75/view?usp=sharing>

Model 2: <https://drive.google.com/file/d/1RkjNsKxqfs1VYEHtF0gIcvqyiUd2vJCu/view?usp=sharing>

Model 3: <https://drive.google.com/file/d/1VqwiCAzBe5N1kWVTkt8te-XfcwVB6ho/view?usp=sharing>

Model 4: https://drive.google.com/file/d/1qe0NNxdbiVH42-1MkooGzKVGxR7_r21a/view?usp=sharing



Figure 85: Valentin Serov - Iphigenia in Tauris

Sonifications:

Model 1: <https://drive.google.com/file/d/1hoaaucwyfw0w9vbi8NcVbUZ31ekpoygu/view?usp=sharing>

Model 2: <https://drive.google.com/file/d/1k5dAGm6NoGk34exzjq80p9qQZJ8y8Vgi/view?usp=sharing>

Model 3: https://drive.google.com/file/d/1xu1FXtP_KA01SiHIS3s-U2q2TM4bSZd5/view?usp=sharing

Model 4: <https://drive.google.com/file/d/1t20LsuMxL4bSb0X0efP9qiDIt6E7pWTc/view?usp=sharing>



Figure 86: Paul Delvaux - The Viaduct

Sonifications:

Model 1: https://drive.google.com/file/d/1YUY1z61NJFj_MMXVN011AjGovGhmP_1R/view?usp=sharing

Model 2: https://drive.google.com/file/d/1swef9SSK0BzDD_aUp44zD5rmvBxrmZ6/view?usp=sharing

Model 3: https://drive.google.com/file/d/11YSHyoMoCMLvRl_WG1vp4Y1i3KytwUmG/view?usp=sharing

Model 4: https://drive.google.com/file/d/10jxghKE2Gn3D_5MCwP8SNFAqN70tgFC2/view?usp=sharing